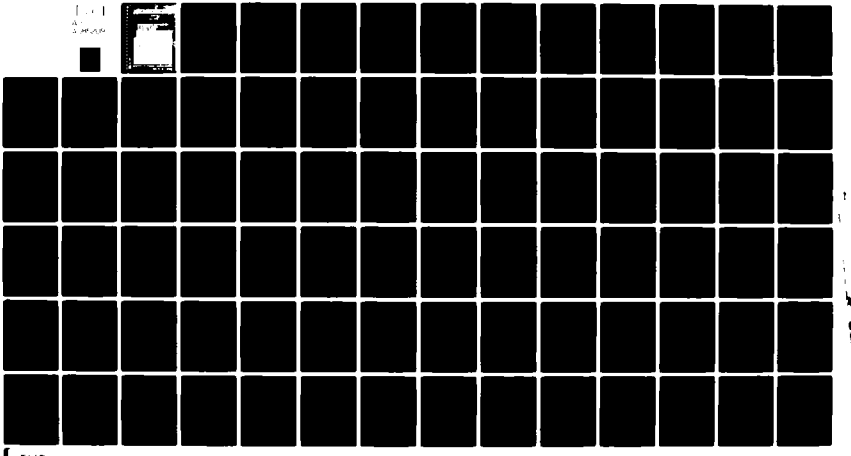
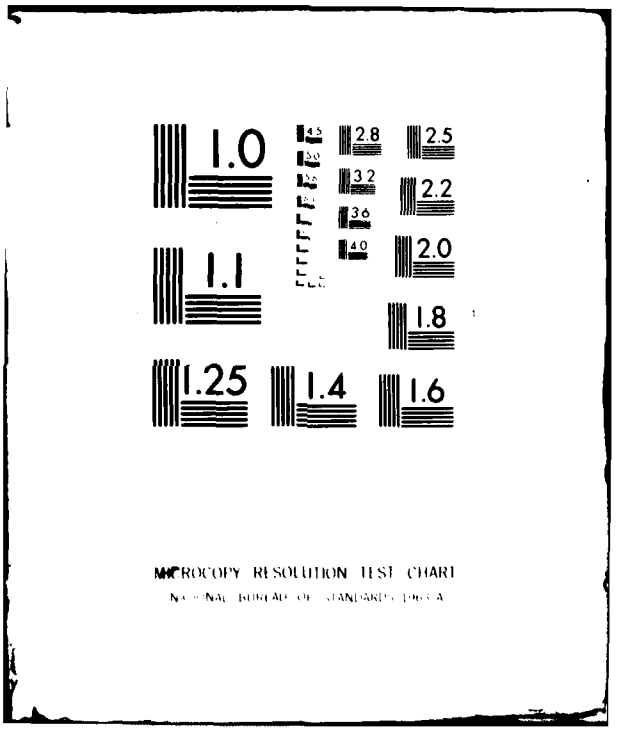
AD-A085 209 ILLINOIS UNIV AT URBANA-CHAMPAIGN COORDINATED SCIENCE LAB F/G 9/3
RANDOMIZATION AND DITHERING IN QUANTIZED SIGNAL DETECTION SYSTE--ETC(U)
OCT 79 M W OAKES N00014-79-C-0424
UNCLASSIFIED R-860 NL



END
DATE FILMED
6 80
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A085209

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A085209 / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) RANDOMIZATION AND DITHERING IN QUANTIZED SIGNAL DETECTION SYSTEMS.	5. TYPE OF REPORT & PERIOD COVERED Technical Report
(14)	6. PERFORMING ORG. REPORT NUMBER R-869, UILU-ENG-78-2253
7. AUTHOR(s) (10) Michael Willard Oakes	8. CONTRACT OR GRANT NUMBER(s) (15) N00014-79-C-0424 / DAAG29-78-C-0016
9. PERFORMING ORGANIZATION NAME AND ADDRESS Coordinated Science Laboratory / University of Illinois / Urbana, Illinois 61801	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Joint Services Electronics Program (11)	12. REPORT DATE Oct[illegible]79 / 13. NUMBER OF PAGES 75
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) (12) 82	15. SECURITY CLASS. (of this report) UNCLASSIFIED / 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

(9) Masters thesis

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Signal Detection, Quantization, Dithering

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Several versions of data quantizers are considered as detection nonlinearities in a binary signal detection problem. Randomized and dithered versions of these quantizers are formulated and the performance of all proposed systems is compared to that of systems using some well-known detection nonlinearities. In particular, the robustness of the quantizer systems in noise with uncertain statistics is compared to that of the well-known systems via asymptotic relative efficiency.

097700 Dm

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

UILU-ENG 78-2253

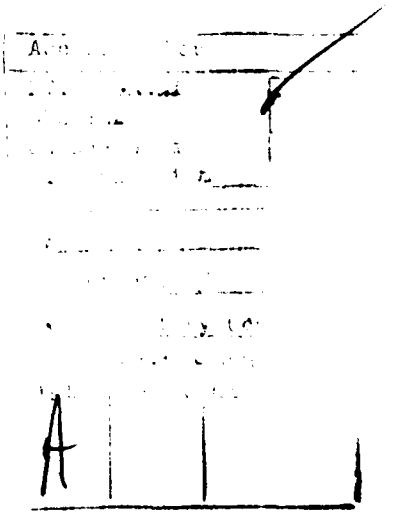
RANDOMIZATION AND DITHERING IN QUANTIZED SIGNAL DETECTION SYSTEMS

by

Michael Willard Oakes

This work was supported by the Joint Services Electronics Program (U.S. Army, U. S. Navy and U. S. Air Force) under Contract DAAG-29-78-C-0016.

Reproduction in whole or in part is permitted for any purpose of the United States Government.



Approved for public release. Distribution unlimited

RANDOMIZATION AND DITHERING IN QUANTIZED SIGNAL DETECTION SYSTEMS

by

Michael W. Oakes
Coordinated Science Laboratory and
Department of Electrical Engineering
University of Illinois at Urbana-Champaign, 1979

ABSTRACT

Several versions of data quantizers are considered as detection nonlinearities in a binary signal detection problem. Randomized and dithered versions of these quantizers are formulated and the performance of all proposed systems is compared to that of systems using some well-known detection nonlinearities. In particular, the robustness of the quantizer systems in noise with uncertain statistics is compared to that of the well-known systems via asymptotic relative efficiency.

RANDOMIZATION AND DITHERING IN QUANTIZED SIGNAL DETECTION SYSTEMS

BY

MICHAEL WILLARD OAKES

B.S., Worcester Polytechnic Institute, 1977

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1980

Thesis Advisor: H. Vincent Poor

Urbana, Illinois

ACKNOWLEDGEMENT

I would like to thank Dr. Poor for his help, advice and suggestions in the execution of this thesis.

TABLE OF CONTENTS

1. INTRODUCTION

Robustness, as the term is used here, refers to the ability of a detection system to perform well despite slight variations in the noise statistics from those used to design the system. In general, for a set of given noise statistics, a robust detector will be outperformed by the optimal detector. For variations in these statistics however, the optimum detector often performs poorly when compared to a robust detector. In this thesis we propose some detectors involving randomized data quantization and dithering and consider their robustness properties relative to other commonly used systems.

The signal detection model we consider here is a simple binary hypothesis test between a hypothesis and an alternative given as follows:

$$H_1: x_i = n_i + \theta s_i$$

$$H_0: x_i = n_i$$

Here, x_i is the i-th observation sample, n_i is noise with a probability density function $f(x)$, s_i is a known signal and θ is a signal-strength parameter.

The detector structure to be considered here is illustrated in Fig. 1. To reflect uncertainties in the noise statistics, we use a model proposed by Huber [1] as used by Martin and Schwartz [2] and hereinafter referred to as the mixture model:

Let $f(x)$ be the noise probability density; we assume

$f(x) = (1-\epsilon)\phi(x) + \epsilon h(x)$; where $0 \leq \epsilon < 1$; (1.1)

$\phi(\cdot)$ is the unit normal density;

$h(\cdot)$ is an arbitrary density; and

ϵ is known and is usually small.

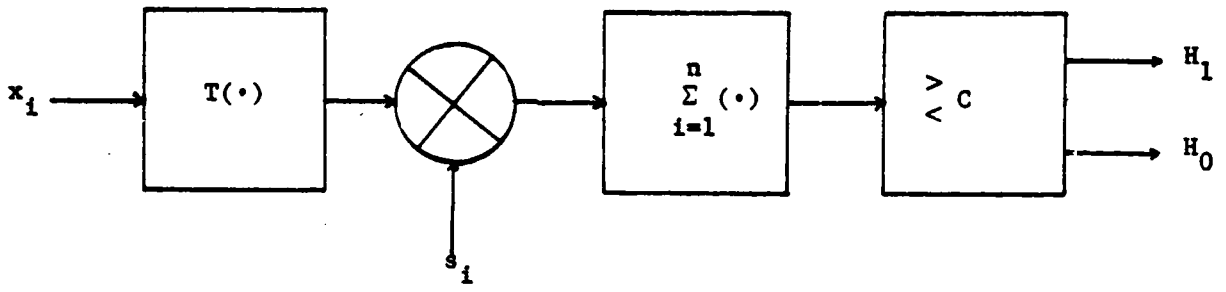


Fig. 1. Detector structure.

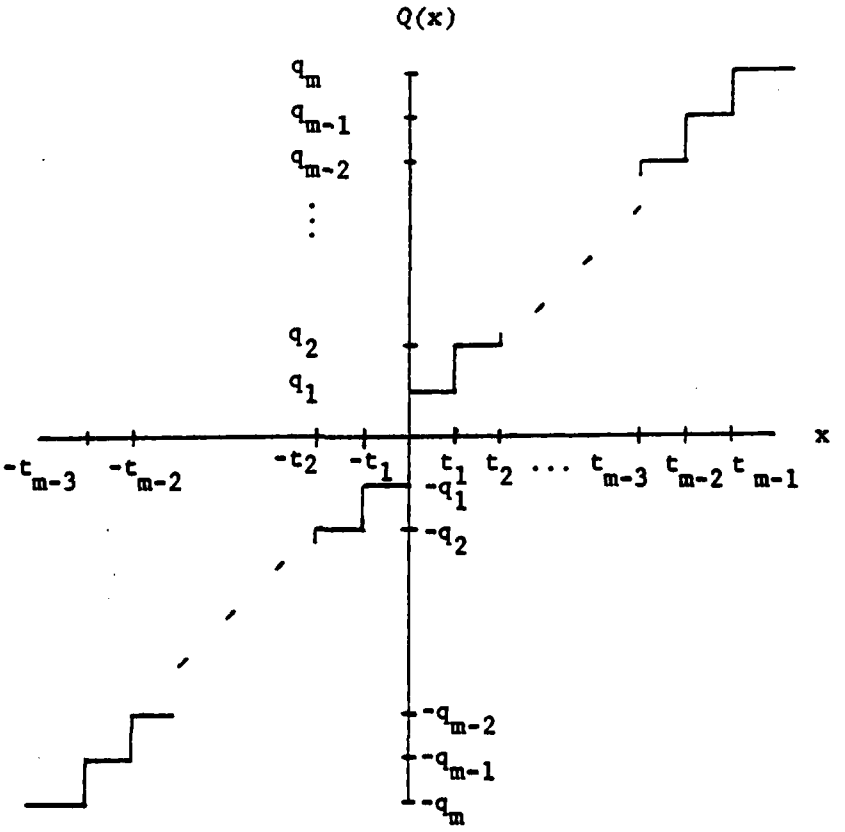


Fig. 2. Typical quantizer.

The clearest comparison of the two systems would be through the power and false alarm rates or the probability of error. This method quickly becomes intractable for the problem under consideration and is abandoned in favor of the asymptotic relative efficiency.

The efficacy of a System A of the form of Fig. 1, is defined as

$$E_A = \lim_{n \to \infty} \frac{\left\{ \frac{\partial}{\partial \theta} E\left[\sum_{i=1}^{n} s_i \, T(x_i) | H_1 \right] \Big|_{\theta=0} \right\}^2}{n \, \mathrm{Var}\left[\sum_{i=1}^{n} s_i \, T(x_i) | H_0 \right]} \tag{1.2}$$

and the performance of System A compared with a System B also of the form of Fig. 1 is given by the ratio of their efficacies which, via the Pitman-Noether Theorem, yields the asymptotic relative efficiency (ARE) (Capon [3]).

$$\mathrm{ARE}_{A,B} = \frac{E_A}{E_B} \tag{1.3}$$

Physically, the ARE represents the savings in samples required by System A to achieve the same power and false alarm rate as System B. Thus if A is more effective than B, $\mathrm{ARE}_{A,B} > 1$. If System A is more complicated than B, then the ARE should be considerably larger than unity to be practically useful.

A quantizer is a nonlinear operation which maps the real numbers to the real numbers in the following manner:

The real line is divided into discrete non-overlapping adjacent segments. A sample point falling in one of these intervals is mapped to a discrete point corresponding to that interval. Let $t_k, k = 1,\ldots,m$ be the set of endpoints of all the intervals $[t_{k-1}, t_k)$ (hereinafter called "breakpoints"). Then for $t_i > 0 \; \forall i$ and $t_0 \triangleq 0$, the positive real numbers are formed by $\mathbb{R}^+ = \lim_{t_m \to \infty} \bigcup_{k=1}^{m} [t_{k-1}, t_k)$. Let the point to which each interval maps be $q_k \geq 0$ (the "level") then the action of the quantizer Q on sample x is $Q(x \mid t_{k-1} \leq x < t_k) = q_k$.

Although not strictly required, in this paper the condition $q_k > q_{k-1}$; $k = 1,\ldots,m$ holds and, in some cases, $q_0 \triangleq 0$. Also, with the exception of the dead-zone detector, $q_k \in [t_k, t_{k-1}]$. Figure 2 shows a typical quantizer.

Examples of quantizer detection systems with fewer restrictions are available, see Poor and Thomas [4,5] or Kassam [6] in particular. In addition, some work has been done to optimize a quantizer with a given number of levels by proper breakpoint and level selection according to some fidelity criterion. See Max [7] or Kassam [6].

Two <u>ad hoc</u> schemes for modifying quantizers are attempted here - randomization and dithering. These techniques are inspired by reported improvement in low-bit digital video by similar methods in two papers, one by Roberts [8] and another by Thompson and Sparkes [9].

Randomization refers to the perturbation from a fixed position of the quantizer breakpoint. This is achieved in the following manner:

One treats the interval length $(t_k - t_{k-1})$ as a random variable which takes a value at each sample independently of its value in other samples. The length of each interval $(t_k - t_{k-1})$ is given the value of this random variable. Appendix B clarifies this approach and gives the derivation of the expression for the efficacy of this type of quantizer-detector (Eq. (B.4)).

Dithering is performed exactly as Roberts performed it with video. A random variable uniformly distributed between $(-t_1/2, t_1/2)$, sample-wise independent is added to the quantizer input and subtracted from the output. Figure 3 clarifies this process. Appendix C contains a derivation of the efficacy for this system.

Appendix A is a derivation of the efficacy of a nonrandomized, nondithered quantizer. A comparison of (A.9), (B.4) and (C.12) will show that the quantizer output levels have no bearing on system performance (as we define it here) except in the dithered system.

In the following sections, specific quantizers and modification methods are described and their performances compared via Eqs. (A.9), (B.4), (C.12), and (1.3).

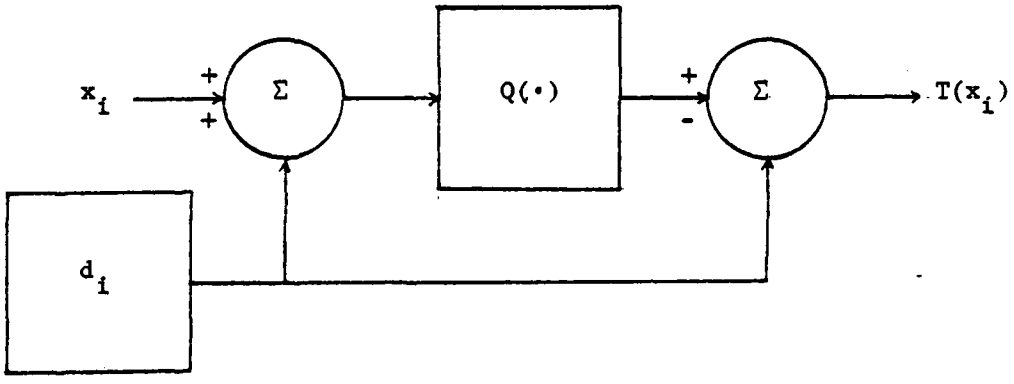


Fig. 3. Dithered quantizer.

2. THE FOUR-LEVEL SYMMETRICAL QUANTIZER

This quantizer is depicted in Fig. 4 and is the simplest quantizer other than a sign detector type or dead-zone quantizer which is relatively easy to optimize. Only three breakpoints are present at 0 and $\pm$ t. The four levels are spaced at $\pm$ q and $\pm$ 2q.

The efficacy of this quantizer when used in the detector structure of Fig. 1 is given by

$$E_{4Q} = \lim_{n \to \infty} \left(\frac{1}{n} \sum_{i=1}^{n} s_i^2 \right) \frac{2}{3} \frac{[f(0) + f(t)]^2}{7/6 - F(t)} \tag{2.1}$$

Where f(x) and F(x) are the density function and distribution function respectively, of the noise.

If s_i *is assumed equal to one for all i (the parameter* θ *will control* the actual received signal strength), then the efficacy E_{4Q} may be normalized to the following:

$$\eta_{4Q} = \frac{2}{3} \frac{[f(0) + f(t)]^2}{7/6 - F(t)} \tag{2.2}$$

The optimum breakpoint t_{opt} can be found by maximizing η_{4Q} over t. This leads to the following necessary condition on t_{opt}:

$$2 \frac{f'(t_{opt})}{f(t_{opt})} \left[\frac{7}{6} - F(t_{opt})\right] + f(0) + f(t_{opt}) = 0 \tag{2.3}$$

where $f'(t_{opt}) \triangleq \frac{d}{dx} f(x)\Big|_{x=t_{opt}}$.

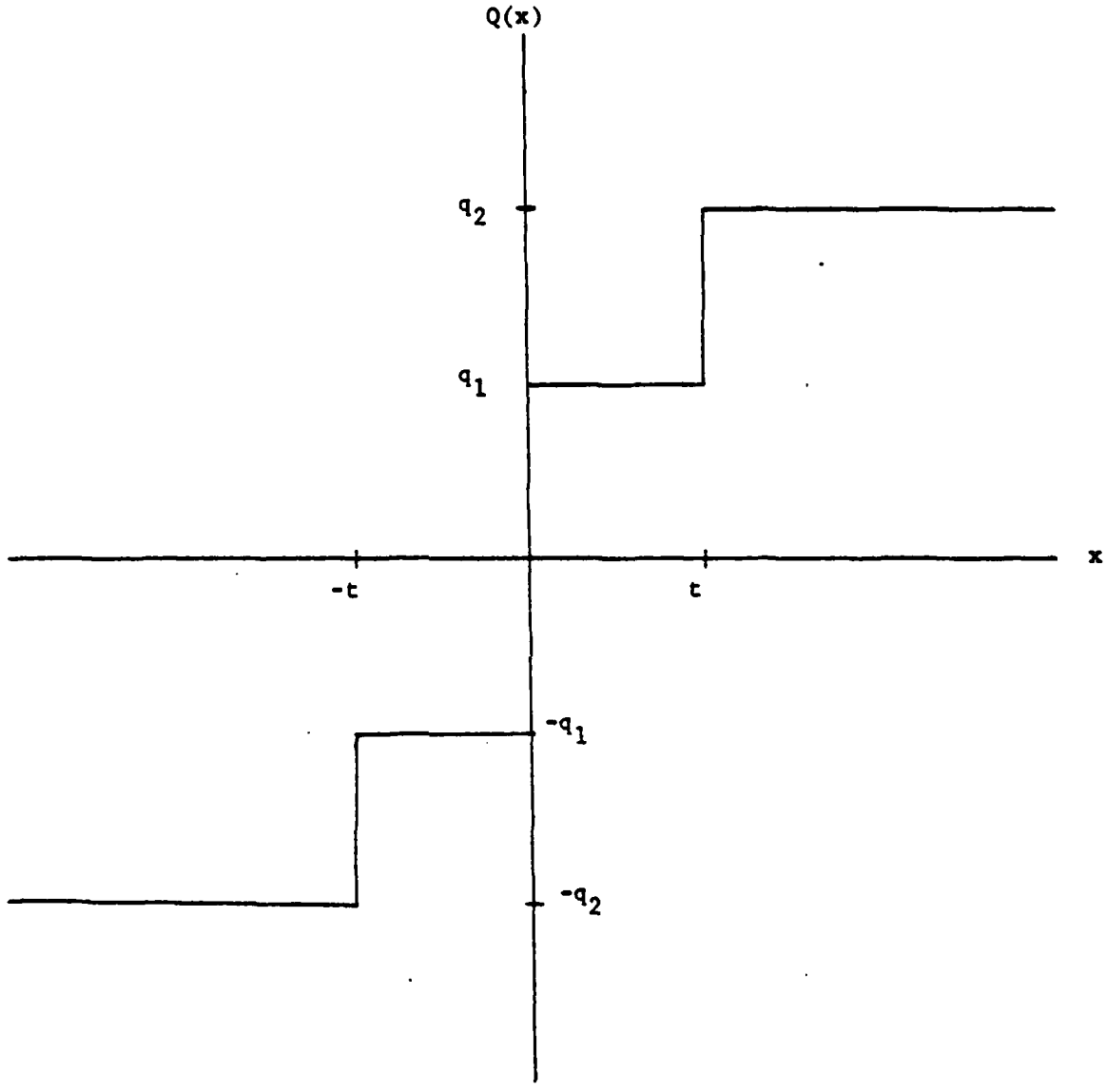


Fig. 4. Four-level symmetrical quantizer.

If $f(x) = \exp(-x^2/2\sigma^2)/(\sqrt{2\pi}\,\sigma)$, Eq. (2.3) is a transcendental equation not directly solvable for t_{opt}. Computer results were used however to produce the estimate $t_{opt} \cong \sigma$ for this case.

Randomization of this quantizer is achieved by making t a random variable with a one sided density over $[0,\infty)$ called g(x) and corresponding distribution function G(x). Then the efficacy is given by (2.4), a special case of (B.4).

$$\eta_{R4Q} = \frac{2}{3}\,\frac{\left[f(0) + \int_0^\infty f(t)dG(t)\right]^2}{7/6 - \int_0^\infty F(t)dG(t)} \tag{2.4}$$

We consider in particular the case where g(x) is a Rayleigh function with parameter $\alpha > 0$:

$$(x) = (x/\alpha^2)\exp(-x^2/2\alpha^2)$$

With f(x) again a Gaussian function, η_{R4Q} becomes

$$\eta_{R4Q} = \frac{2}{\pi}\,\frac{(\alpha^2 + 2\sigma^2)^2}{\sigma^2[4(\alpha^2 + \sigma^2)^2 - 3\alpha(\alpha^2 + \sigma^2)^{3/2}]} \tag{2.5}$$

This can be maximized over α to find α_{opt}.

$$\alpha_{opt} = \sigma\sqrt{K}\ ,\ K \cong .4063367$$

This yields $\eta_{R4Q}\big|_{\alpha=\alpha_{opt}} \cong .7807002/\sigma^2$

The performances of the two quantizers were compared using the mixture model of (1.1) with h(x) a Gaussian density with variance σ^2. The results of this comparison are shown graphically in Fig. 5. It is seen that for large contamination the randomized quantizer performs only slightly better. The trend for small values of σ seems to indicate a large performance improvement but the small range over which this occurs may not compensate for the extra complexity of the system necessary to achieve this improvement.

It is worth noting again that the actual level value, q, does not appear in the efficacy expression. The output of the quantizer indicates into which interval the received sample falls. This information alone is all that is necessary to make a decision between H_0 and H_1, therefore the actual value of the level is arbitrary.

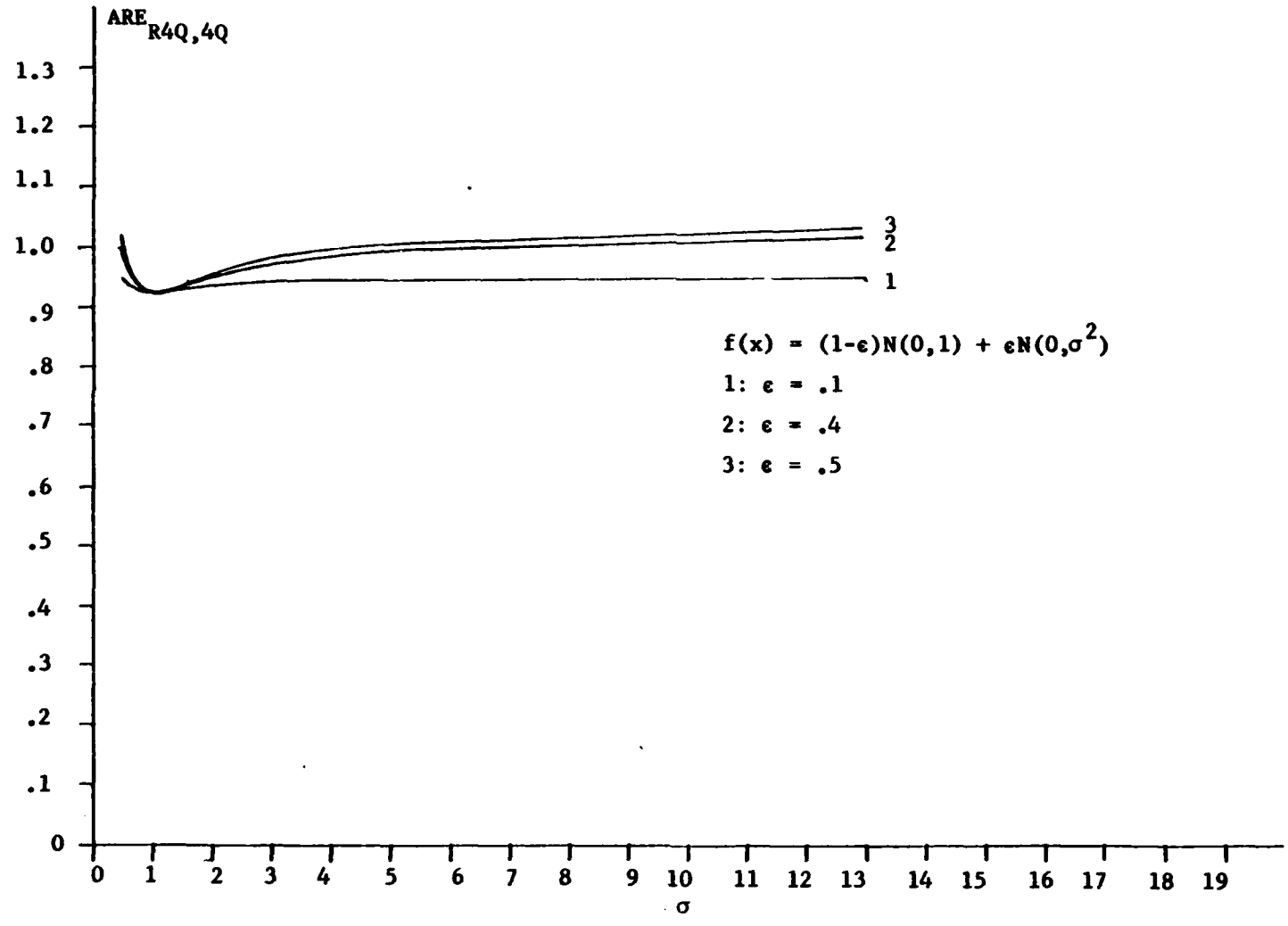


Fig. 5. Four-level quantizer performance.

3. THE DEAD-ZONE QUANTIZER

The dead-zone or null-zone quantizer, as taken here, is a three level symmetrical quantizer with output levels 0, $\pm$ q and breakpoints $\pm$ t. This type of detector was studied in a different context by Kassam and Thomas [10]. Their use of it is different enough, however, that the results are not comparable with those given here. Figure 6 illustrates the dead-zone quantizer.

The efficacy of this quantizer is given by (3.1) for a symmetrical noise density function f(x).

$$\eta_{DZ} = \frac{2f^2(t)}{1 - F(t)} \tag{3.1}$$

An optimum breakpoint t_{opt} can be found for a given density f by maximizing η_{DZ} over t. If f is Gaussian with variance σ^2, then $t_{opt} = \sigma$.

Randomization is achieved here by letting the breakpoint t be a random variable with some distribution G(x) (and corresponding density g(x)). The efficacy of the randomized dead-zone quantizer then becomes:

$$\eta_{RDZ} = \frac{2\left[\int_0^\infty f(t)dG(t)\right]^2}{\int_0^\infty [1-F(t)]dG(t)} \tag{3.2}$$

Three possible density functions were examined, each with a single parameter which could be adjusted to maximize (3.2). These are listed below

1. Single-Sided Exponential

$$g(x) = \gamma \exp[-\gamma x] \cdot u(x) \tag{3.3}$$

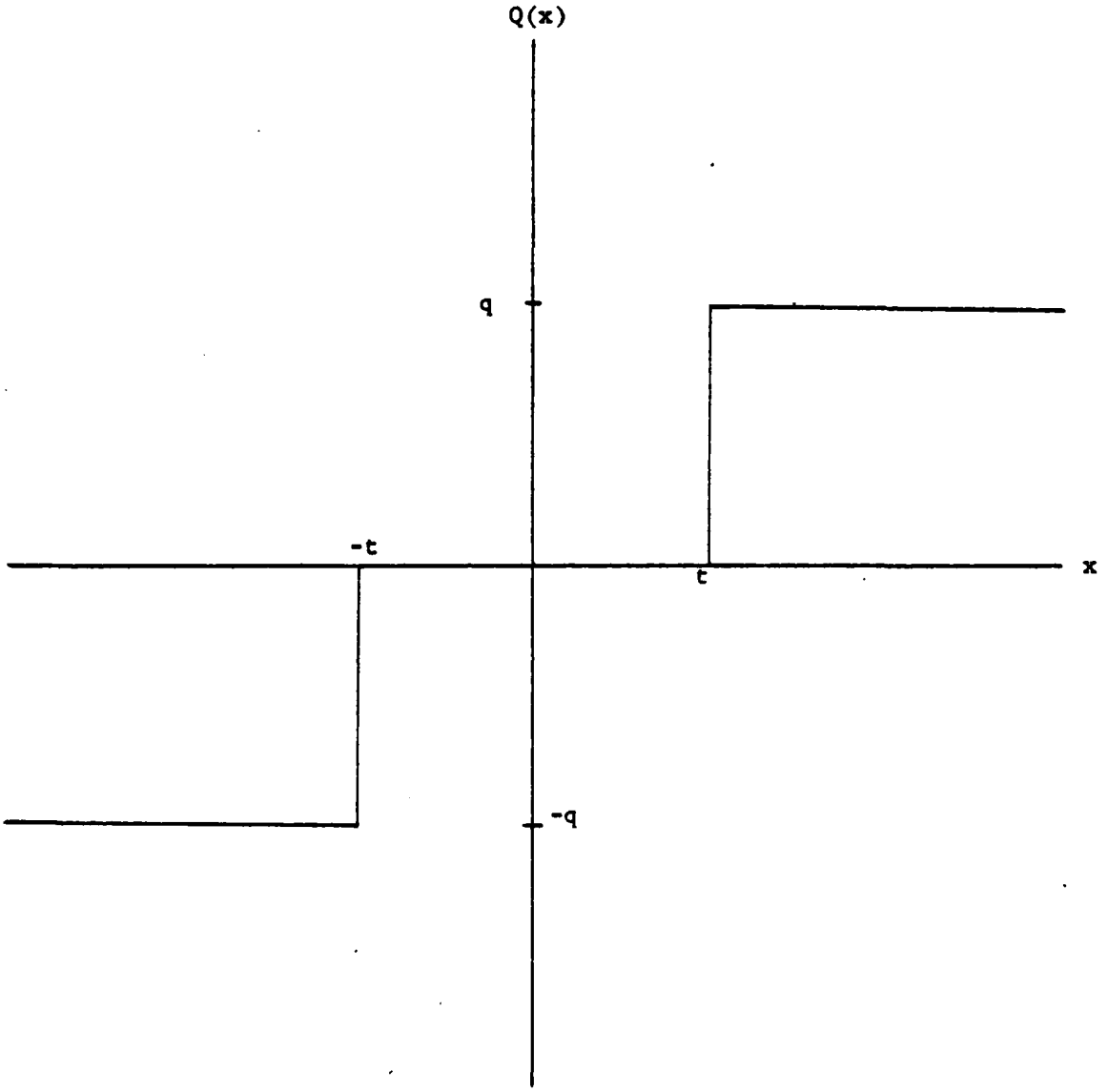


Fig. 6. Dead-zone quantizer.

2. Rayleigh

$$g(x) = (x/\alpha^2)\exp[-x^2/(2\alpha^2)]\cdot u(x) \tag{3.4}$$

3. Triangular

$$g(x) = \begin{cases} x/r^2 & 0 \le x \le r \\ -x/r^2 + 2/r & r < x \le 2r \\ 0 & \text{elsewhere} \end{cases} \tag{3.5}$$

Since the nominal noise in the mixture model of (1.1) used throughout this paper is unit variance Gaussian noise, the three parameters were each optimized for f(x) a Gaussian with variance σ^2.

Under these conditions then, the following results are achieved:

$$\begin{aligned} \gamma_{opt} &= +\infty \\ \alpha_{opt} &= \sigma/2\sqrt{2} \\ r_{opt} &= 0 \end{aligned} \tag{3.6}$$

Clearly γ_{opt} and r_{opt} lead to degenerate forms in (3.3) and (3.5). An arbitrary value of γ was used to get some results but the use of the triangular density was abandoned at this point. Performance measurements are based upon an exponential contamination noise, that is, in reference to the previously described mixture model, we have

$$h(x) = (\beta/2)\exp[-\beta|x|], \ \beta > 0. \tag{3.7}$$

Figure 7 shows the performance of the nonrandomized dead-zone quantizer compared with that of the 4-level quantizer described above. Three values of ϵ (.1, .2 and .3) are used and indicate the trend of the ARE for this case. The dead-zone quantizer performs only slightly better for $\epsilon > .2$ and only in a narrow range of β.

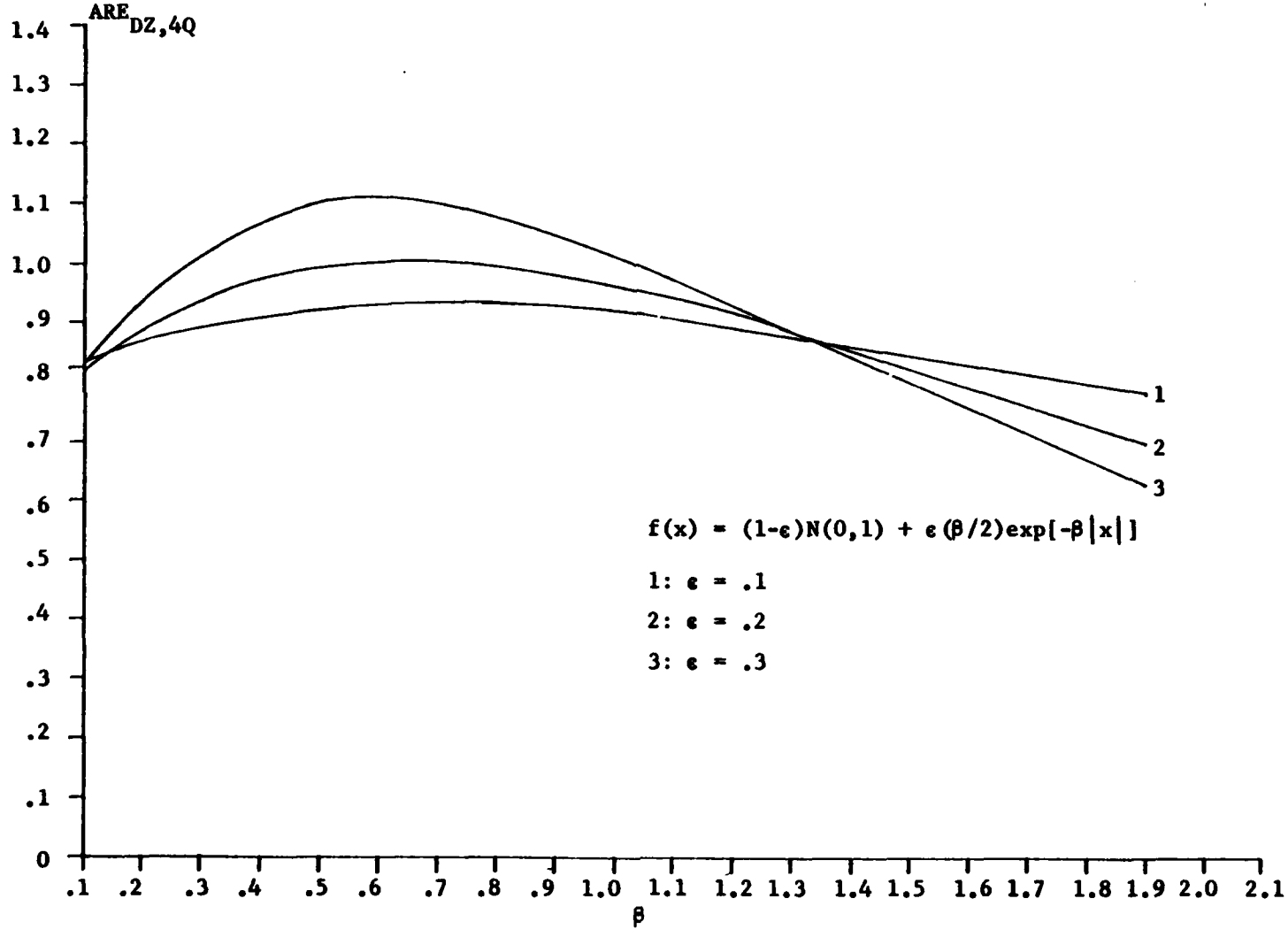


Fig. 7. Dead-zone and four-level quantizer comparison.

The small amount of improvement and the narrow region over which it occurs are reasons enough to exclude the dead-zone quantizer as a replacement for the 4-level quantizer. Coupled with these reasons is the question of the validity of the mixture model as a representation of uncertainty in the noise model when ϵ becomes large. The mixture model here is meant to represent noise that is primarily Gaussian. The detector structure is designed with this assumption and its performance monitored as the noise deviates from unit variance Gaussian, the amount of deviation being represented by ϵ and the "direction" of deviation depending upon the function used for $h(x)$. Thus if ϵ is large ($\epsilon \geq .5$ is probably a liberal definition of large) one would do better to abandon the mixture model and examine some other technique for detector design.

Does randomization improve the performance? Figure 8 shows the ARE of the exponentially randomized dead-zone (RDZE) compared with the nonrandomized version. The value used for γ is $\gamma = 1/\sqrt{2\pi}$. This is an arbitrary value and the results are not good. Since $\gamma = +\infty$ maximizes (3.2), these results are not unexpected.

Figure 9 shows the change in performance from using a Rayleigh distributed breakpoint. Once again, the improvement is slight and occurs only for small β and large ϵ.

Dithering also provides a possible modification of the dead-zone quantizer. This process again is best summarized by Fig. 3. In the application of dithering to this case the random variable d_1 (hereafter called the dither signal) is given a uniform distribution in the interval $[-c,c]$ where c is a positive constant. Two cases are considered here: 1) $c = t$, 2) $c = t/2$.

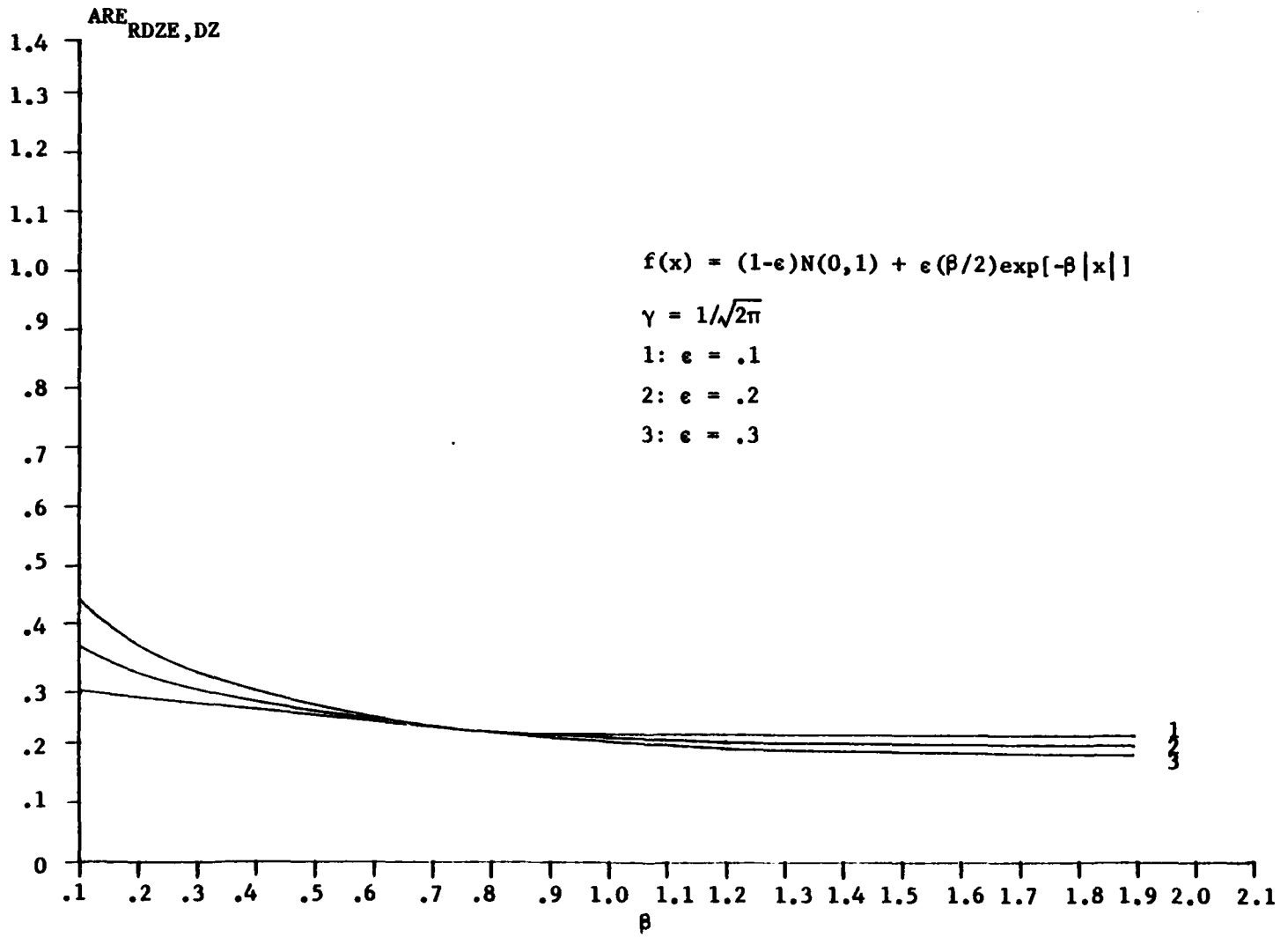


Fig. 8. Exponentially randomized dead-zone compared with dead-zone quantizer.

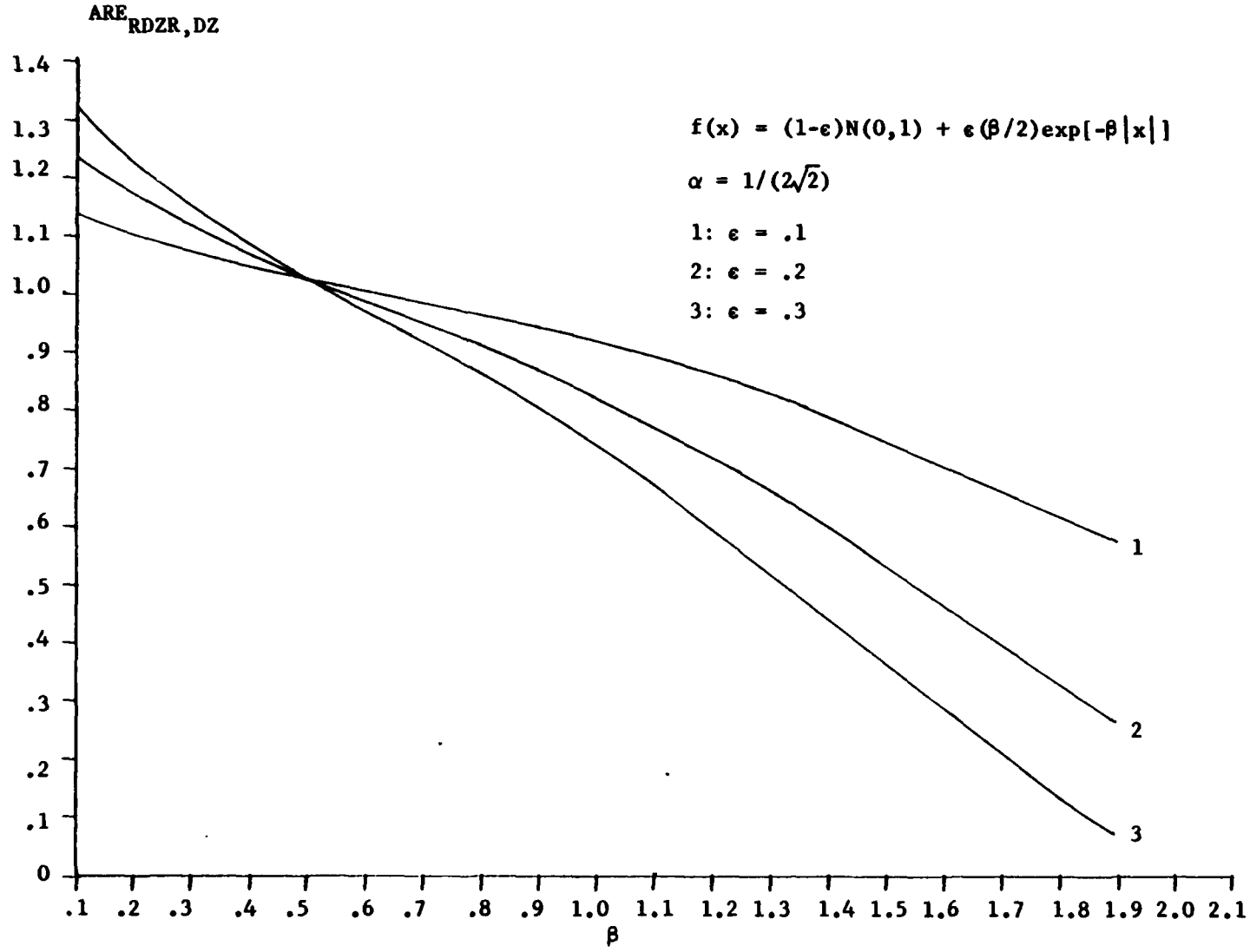


Fig. 9. Rayleigh randomized dead-zone compared with dead-zone quantizer.

Letting $G(x)$ and $g(x)$ be the distribution and density respectively of the dither signal we have the following expressions for the efficacy (3.8) and the normalized efficacy (3.10) of the dithered dead-zone quantizer.

$$E_{DDZ} = \lim_{n\to\infty} \frac{1}{n} \frac{(q\Sigma s_i^2)^2\left[\int_{-\infty}^{\infty}\{f(d_i+c)+f(d_i-c)\}dG(d_i)\right]^2}{q^2\Sigma s_i^2\left[1-\int_{-\infty}^{\infty}\{F(d+c)-F(d-c)\}dG(d_i)\right]-2q\Sigma s_i^2\left[\int_{-\infty}^{\infty}\{f(d_i+c)+f(d_i-c)\}dG(d_i)\right]}$$

$$-\left[q\Sigma s_i\left(1-\int_{-\infty}^{\infty}\{F(d_i+c)+F(d_i-c)\}dG(d_i)\right)\right]^2+\Sigma s_i^2\,\mathrm{Var}[d_i]\Big) \qquad (3.8)$$

where all summations are over i from 1 to n.

Under the assumption that d_i and the noise are sample-wise independent we make the following assumption:*

$$\lim_{n\to\infty} \frac{n(\Sigma\, s_i)^2}{(\Sigma\, s_i^2)^2} = 0 \qquad (3.9)$$

Under (3.9) then, the next to last term in the denominator of (3.8) becomes zero in the limit. Since this term is subtracted, and the term itself is positive, the denominator value will only be increased and at worst, the normalized efficacy will be lower than if the term were maintained. Thus, η_{DDZ} represents a worst case efficacy in those cases where (3.9) does not strictly hold.

$$\eta_{DDZ} \cong \frac{\left[\int_{-\infty}^{\infty}\{f(d+c)+f(d-c)\}dG(d)\right]^2}{1-\int_{-\infty}^{\infty}\{F(d+c)-F(d-c)\}dG(d)-\frac{2}{q}\int_{-\infty}^{\infty}\{f(d+c)+f(d-c)\}dG(d)+\frac{c^2}{3q^2}} \qquad (3.10)$$

*This assumption would be true, for example if $s_i = (-1)^i$; $i = 1,2,\ldots,n$.

Note that dithering has increased the complexity of the efficacy expression and has brought the level value q into play. We may now find a q_{opt} which maximizes the denominator of (3.10).

Using a Gaussian distribution for F(x), q_{opt} was found for Cases 1 and 2:

$$\underline{\text{Case 1}}: \quad c = t, \; q_{opt} = \left(\frac{2c^2}{3}\right)\Big/\left(\frac{2\sigma}{\sqrt{2\pi}} - \frac{\sigma^2}{c}\left\{\Phi\left(\frac{2c}{\sigma}\right) - \frac{1}{2}\right\}\right) \tag{3.11}$$

$$\underline{\text{Case 2}}: \quad c = t/2$$

$$q_{opt} = \left(\frac{c^2}{\sigma}\right)\Big/\left(c + \frac{4}{c}\left[(c^2+\sigma^2)\left\{\Phi\left(\frac{c}{2\sigma}\right) - \Phi\left(\frac{3c}{2\sigma}\right)\right\}\right] + \frac{2\sigma}{\sqrt{2\pi}}\left[\exp\left(\frac{-c^2}{8\sigma^2}\right) - \exp\left(\frac{-9c^2}{8\sigma^2}\right)\right]\right) \tag{3.12}$$

The fact that q enters into these calculations at all is directly attributed to the *dithering process*, therefore one should not be surprised that different densities for the dither signal alter the nature of q_{opt}. In Case 1, it is only necessary for the received signal to vary from 0 for the dither signal to affect the quantizer output whereas in Case 2, the input may be anywhere inside the interval [-t/2,t/2] before the dither comes into play.

The actual derivation of Eqs. (3.11) and (3.12) is tedious but straightforward and will not be repeated here.

A check of the efficacy values for Case 1 and Case 2 using only Gaussian noise (no mixture model) for various values of σ showed that Case 1 was uniformly better. The dead-zone quantizer itself was optimized with $t = \sigma$. Table 1 summarizes the results for a few values of σ.

TABLE 1

Efficacy of Dithered Dead-Zone Quantizer

σ	Case 1	Case 2
.5	2.56	.30
1.0	.73	.09
1.5	.28	.04
2.0	.12	.02
2.5	.05	.02

Based on these results Case 2 was abandoned in favor of Case 1. Henceforth all references to the dithered dead-zone quantizer will imply the use of the Case 1 density for the dither signal.

Results were also obtained for values of t other than $t = \sigma$ and showed that the efficacy is generally better for a value of t slightly less than σ. For this reason, results obtained with the mixture model use $t = .9$.

Both the Gaussian mixture model ($h(x) = N(0,\sigma^2)$) and the exponential mixture model (h(x) as in Eq. (3.7)) were used to evaluate the dithered dead-zone quantizer. Figures 10 through 16 summarize the results . In addition to the comparison with the nondithered, nonrandomized dead-zone, Figs. 12, 15 and 16 show the comparison to the sign detector. These curves serve to relate the performance of the system to a well-known robust detector. Equation (3.13) gives the efficacy of the sign detector.

$$\eta_{SD} = 4 f^2(0) \tag{3.13}$$

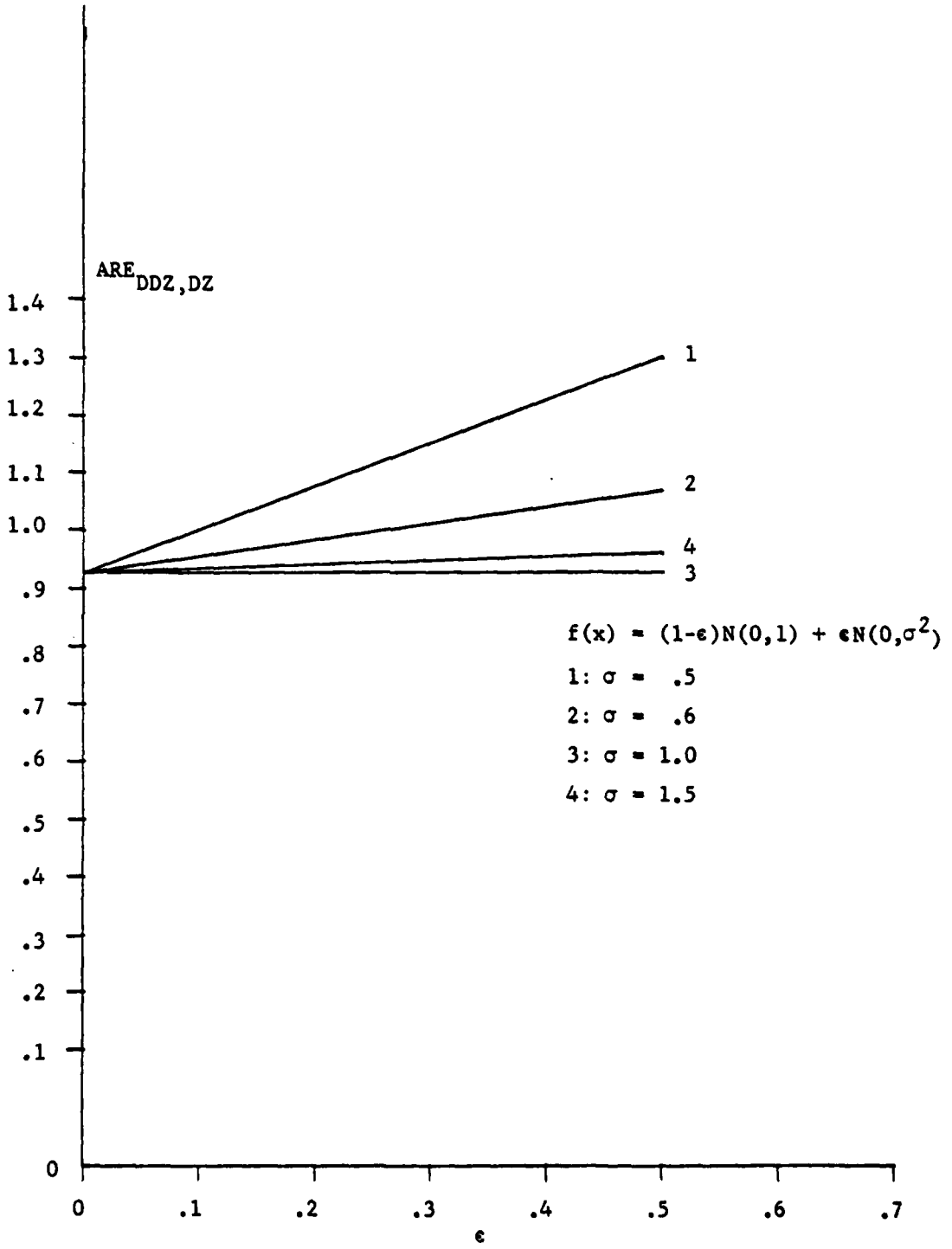


Fig. 10. Dithered dead-zone compared with dead-zone quantizer.

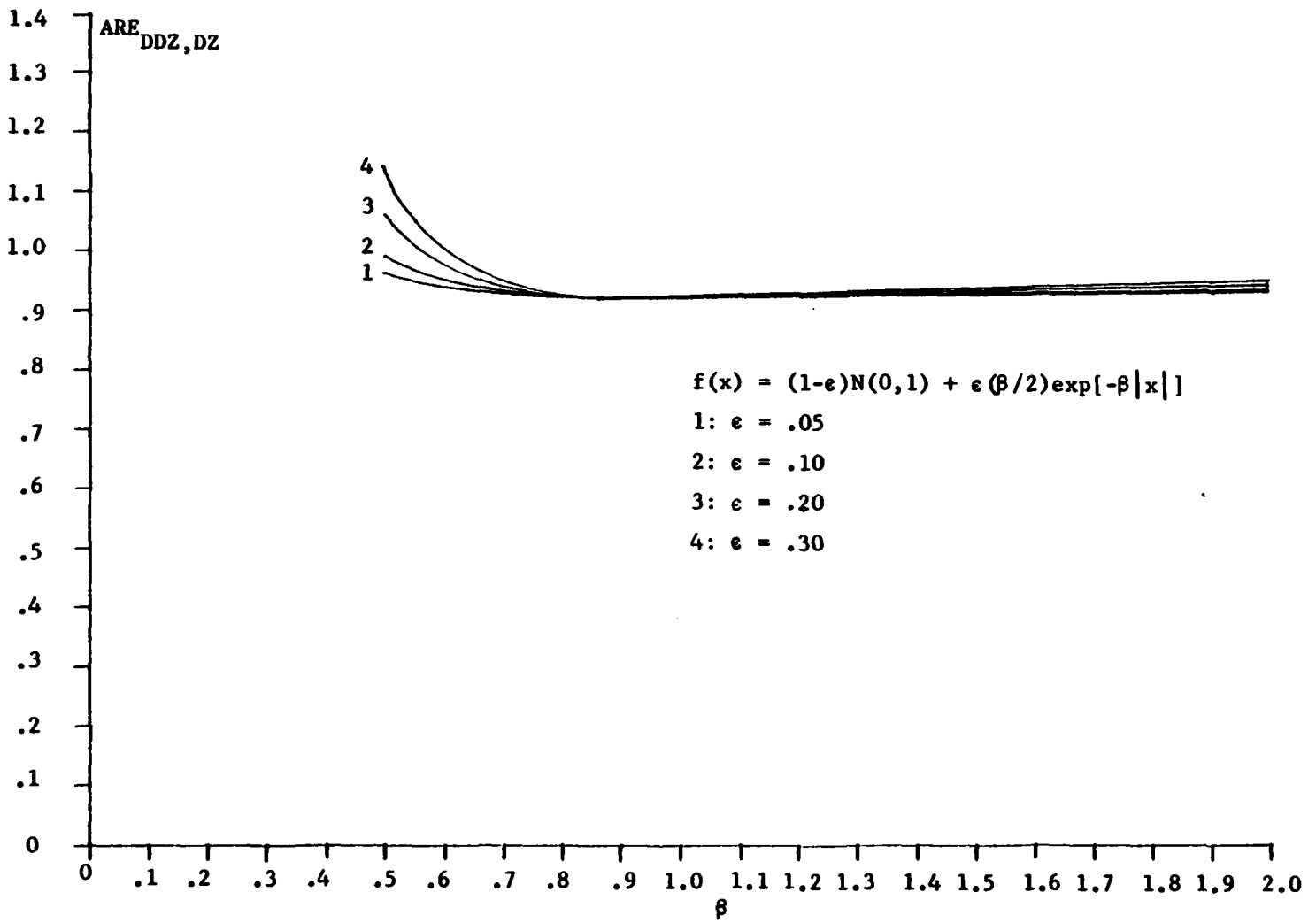


Fig. 11. Dithered dead-zone compared with dead-zone quantizer.

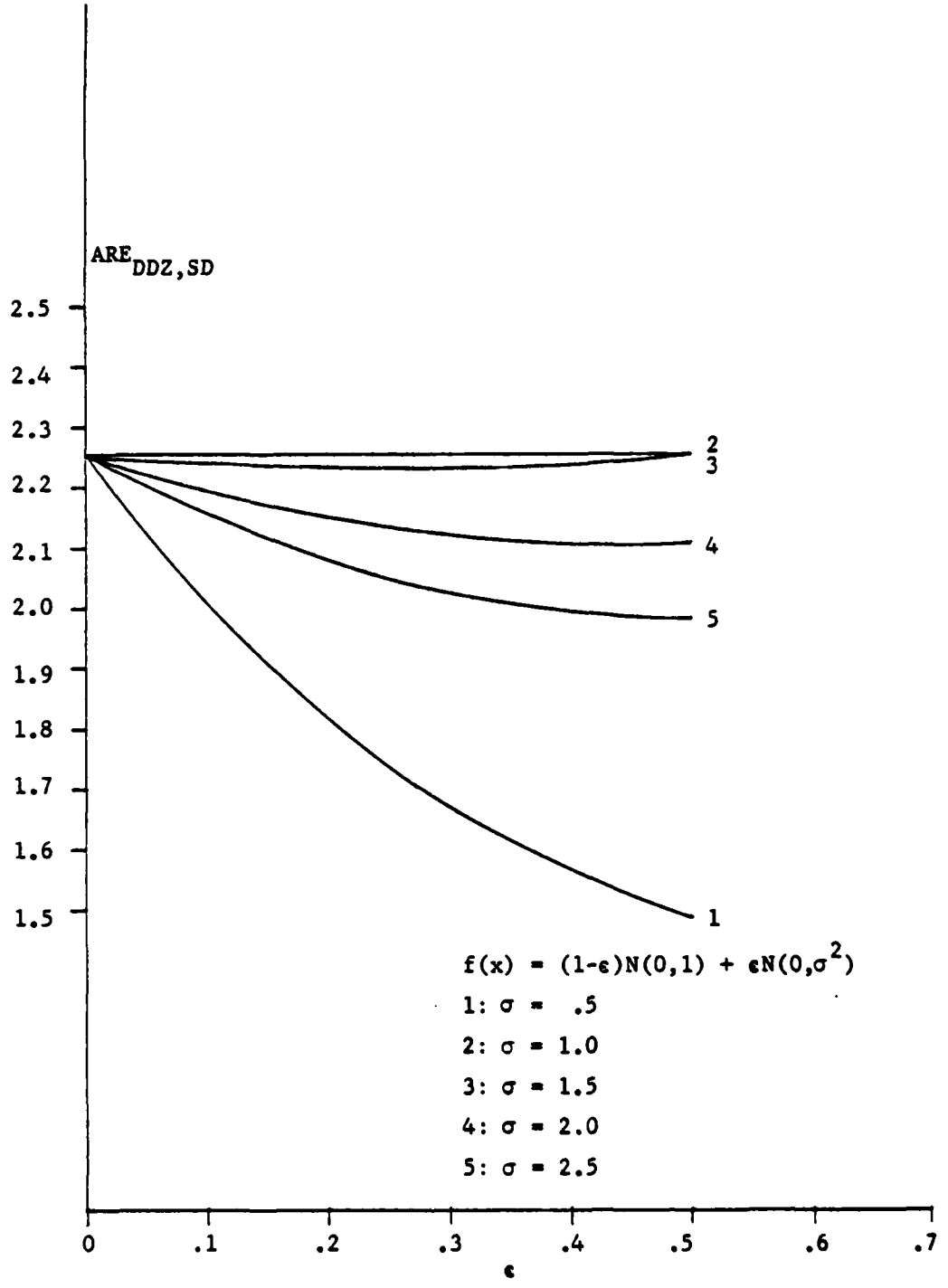


Fig. 12. Dithered dead-zone compared with sign detector.

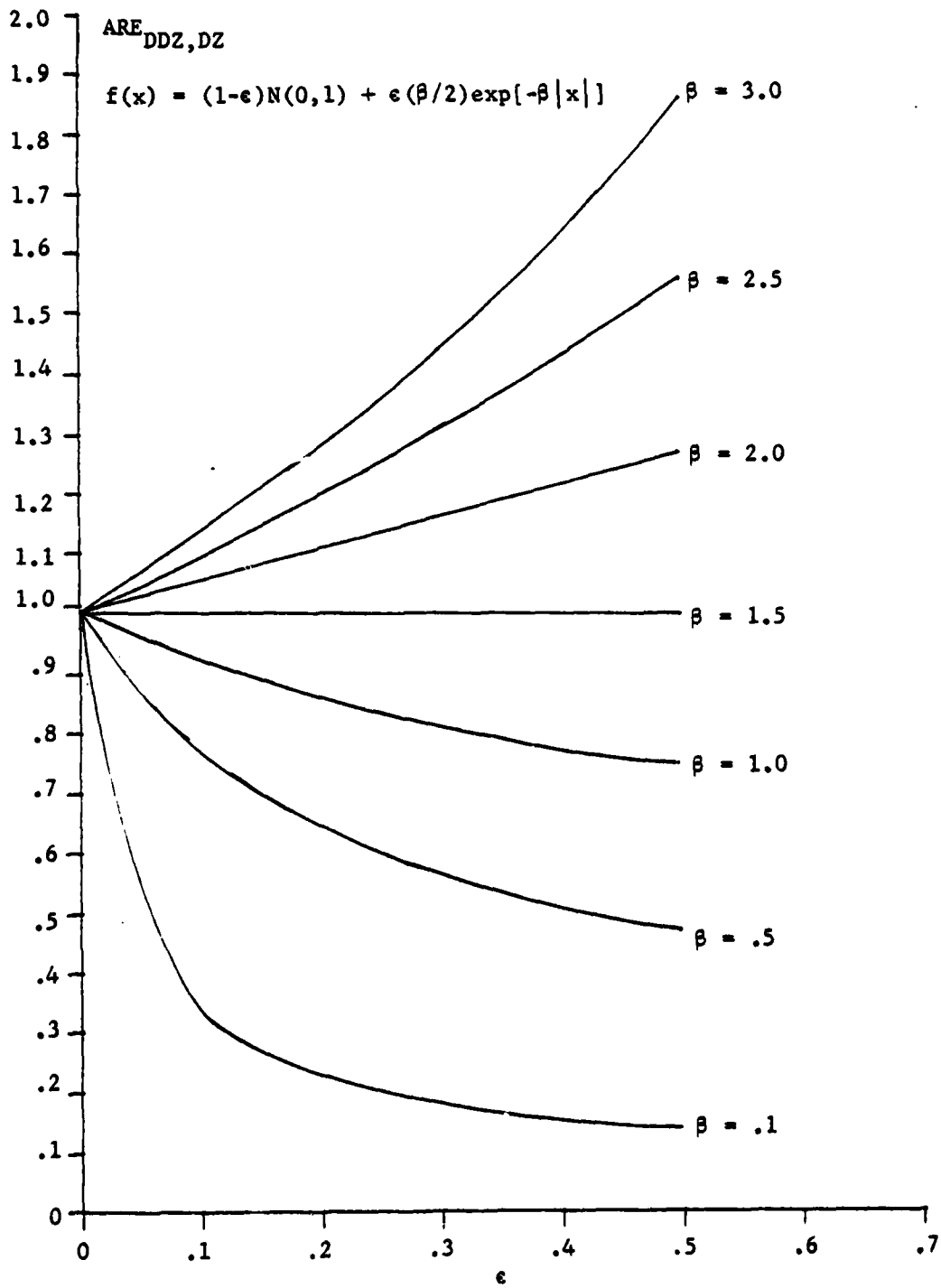


Fig. 13. Dithered dead-zone compared with dead-zone quantizer.

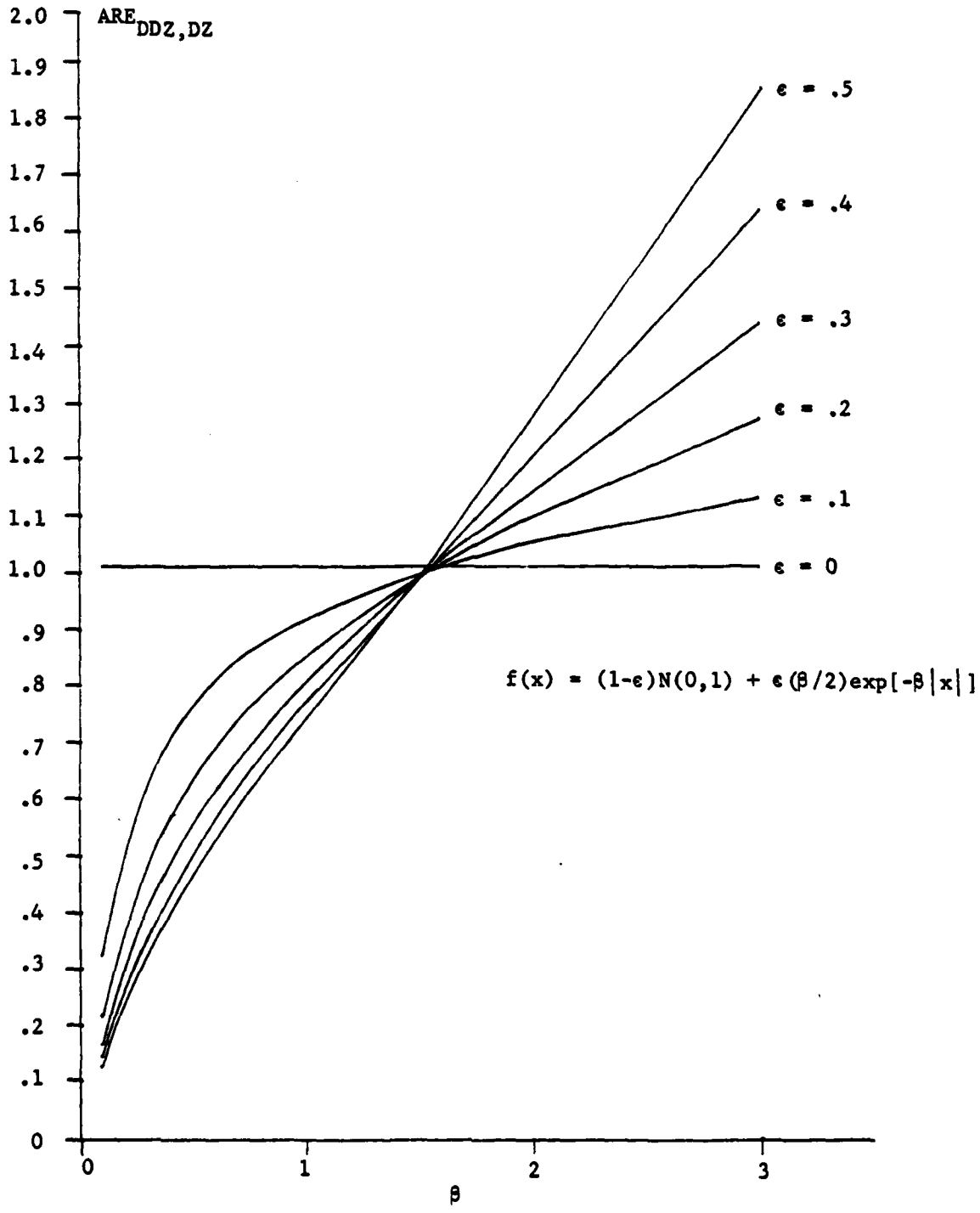


Fig. 14. Dithered dead-zone compared with dead-zone quantizer.

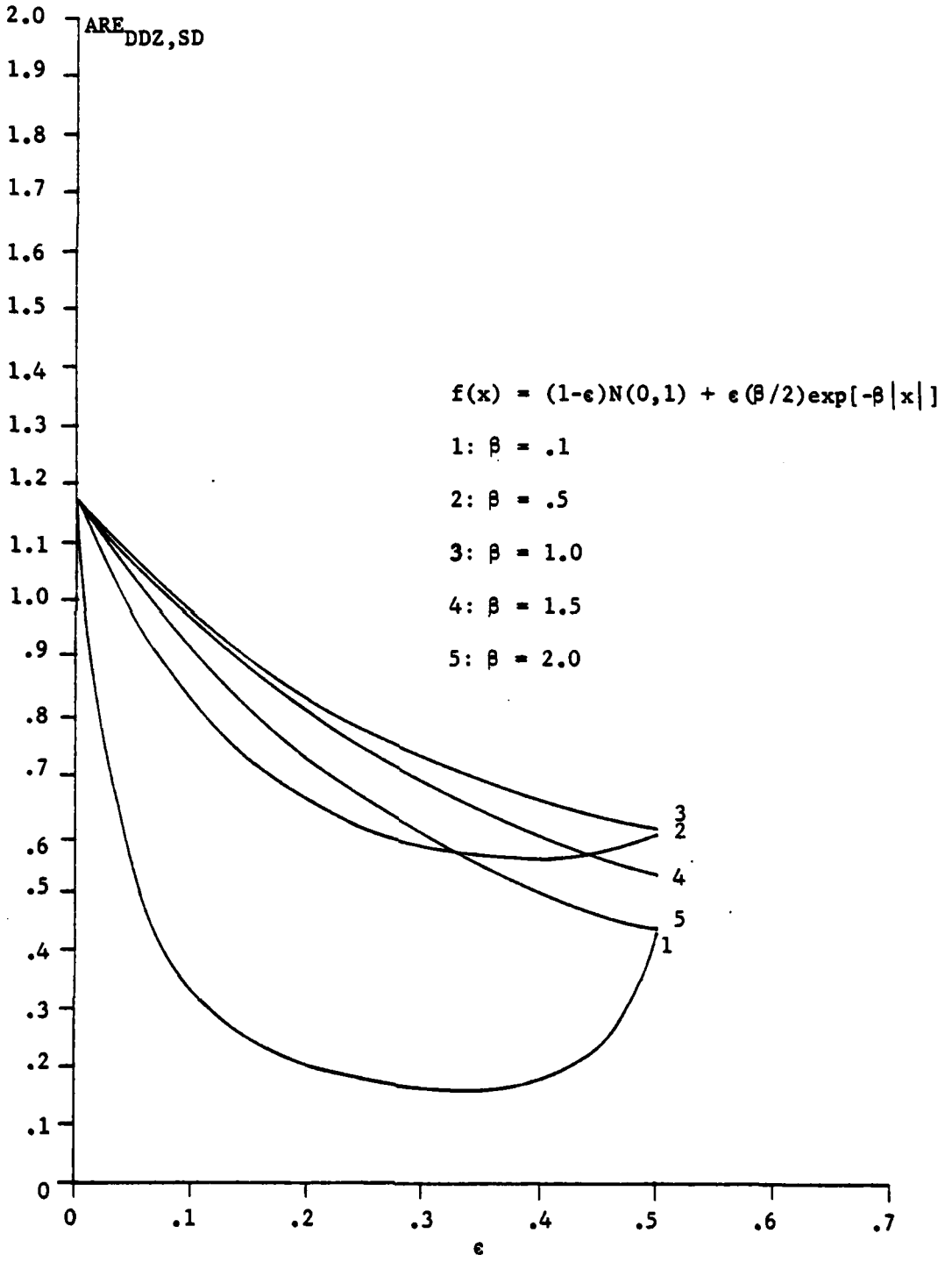


Fig. 15. Dithered dead-zone compared with sign detector.

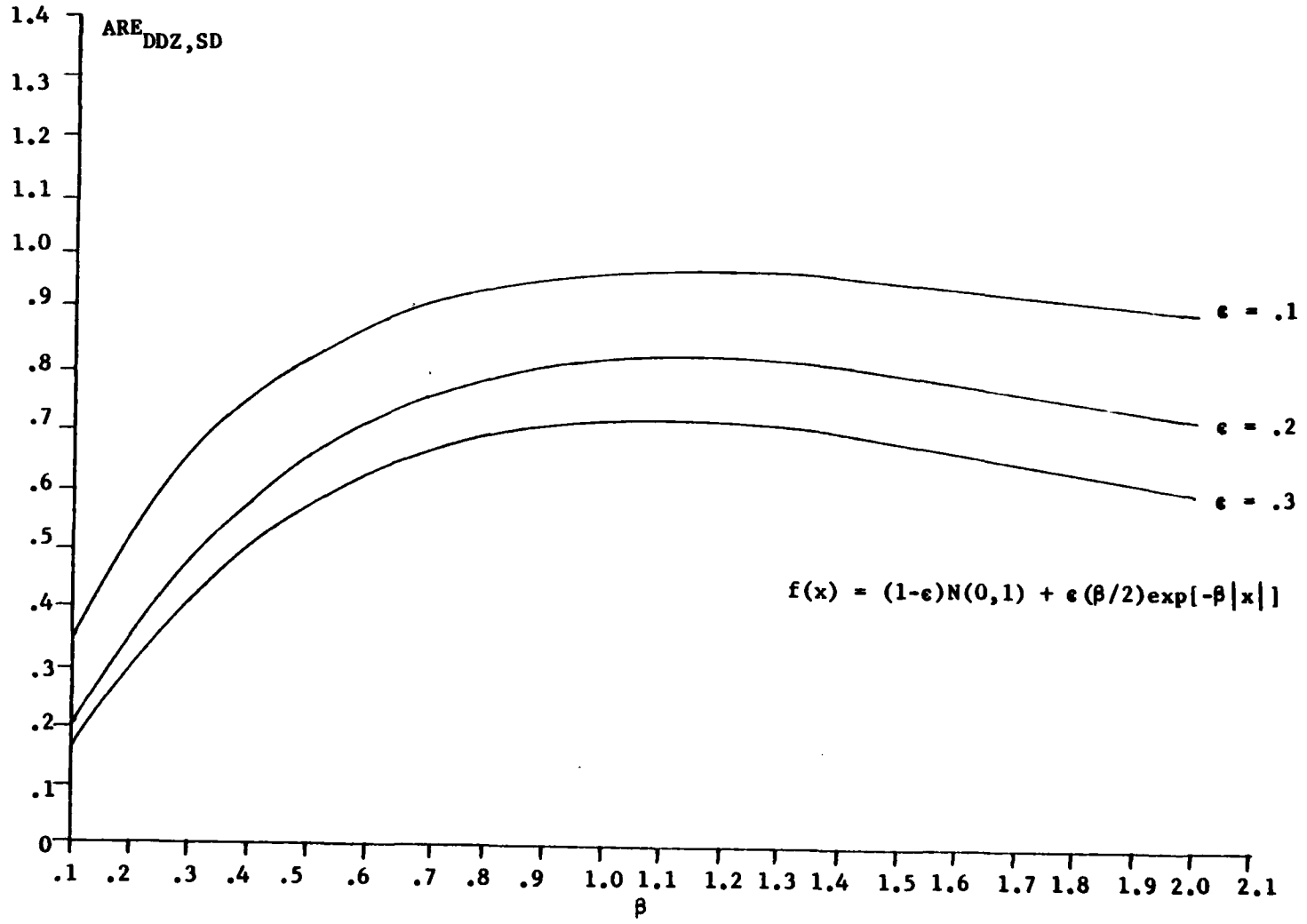


Fig. 16. Dithered dead-zone compared with sign detector.

Comparing Figs. 10 and 13 or Figs. 11 and 14, improvement with dithering is obtained for small values of σ and large values of β. (Note that these refer to two different mixture models.) Both of these cases represent densities h(x) which are becoming more sharply peaked at 0. That is, as σ decreases in the Gaussian mixture model, and as β increases in the exponential mixture model, h(x) becomes more sharply peaked at 0. This improvement is clearly much greater for the exponential contamination however and extends for all $\beta > 1.5$. The improvement for Gaussian contamination is not only less marked but exists only for $\sigma < 1$ and larger values of ϵ which again brings to mind the question of the validity of the results as a basis for detector design.

Comparison of Figs. 12 and 15 or 16 show that, under Gaussian contamination, the dithered system performs considerably better than a sign detector over a wide range of both σ and ϵ, whereas under exponential contamination the dithered is worse by far except for very small ϵ. Even for this small ϵ the $ARE_{DDZ,SD} = 1.2$ which is not a significant improvement if one considers the additional complexity of the dithered system.

In summary, dithering seems to produce unpredictable results. Performance with sharply peaked contamination densities is improved but that improvement is sensitive to the density function and may not be significant when the dithered dead-zone is compared to some other detection system (e.g., the sign detector).

In an attempt to make the performance of the dithered dead-zone less sensitive to the contamination density, a sign nonlinearity can be placed in line following the correlator. For brevity, this new detector is referred to as the dithered sign detector although it is not really a dithered version of the sign detector. Figure 17 will clarify the structure of this system.

Keeping all notation from the preceding discussion of the dead-zone quantizer, we have the efficacy of the dithered sign detector:

$$\eta_{DSD} = \left[\int_{-\infty}^{\infty} \{f(t-d) + f(-t-d)\}dG(d)\right]^2 \qquad (3.14)$$

In order to compare this with the previous system all parameters of the dead-zone quantizer are maintained. That is, G(d) represents a uniform distribution over $[-t,t]$, $t = .9$ and q is given by (3.11).

The results of comparing the dithered sign detector to both the dead-zone quantizer and the sign detector in Gaussian and exponential contaminated noise are presented in Figs. 18 through 21. These graphs show almost uniform degradation in performance. The only improvement appears in the comparison to the sign detector in Fig. 21. It appears here that, for decreasing β and for $\epsilon > .4$, the $ARE_{DSD,SD}$ increases drastically. The importance of this result is once again undermined by the large value and narrow range of ϵ over which this occurs.

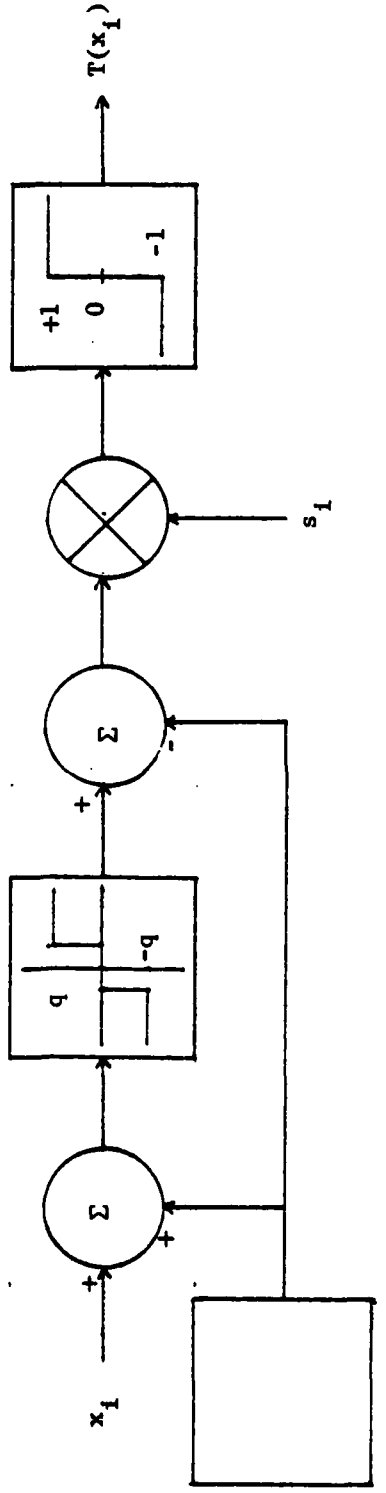


Fig. 17. Dithered Sign Detector.

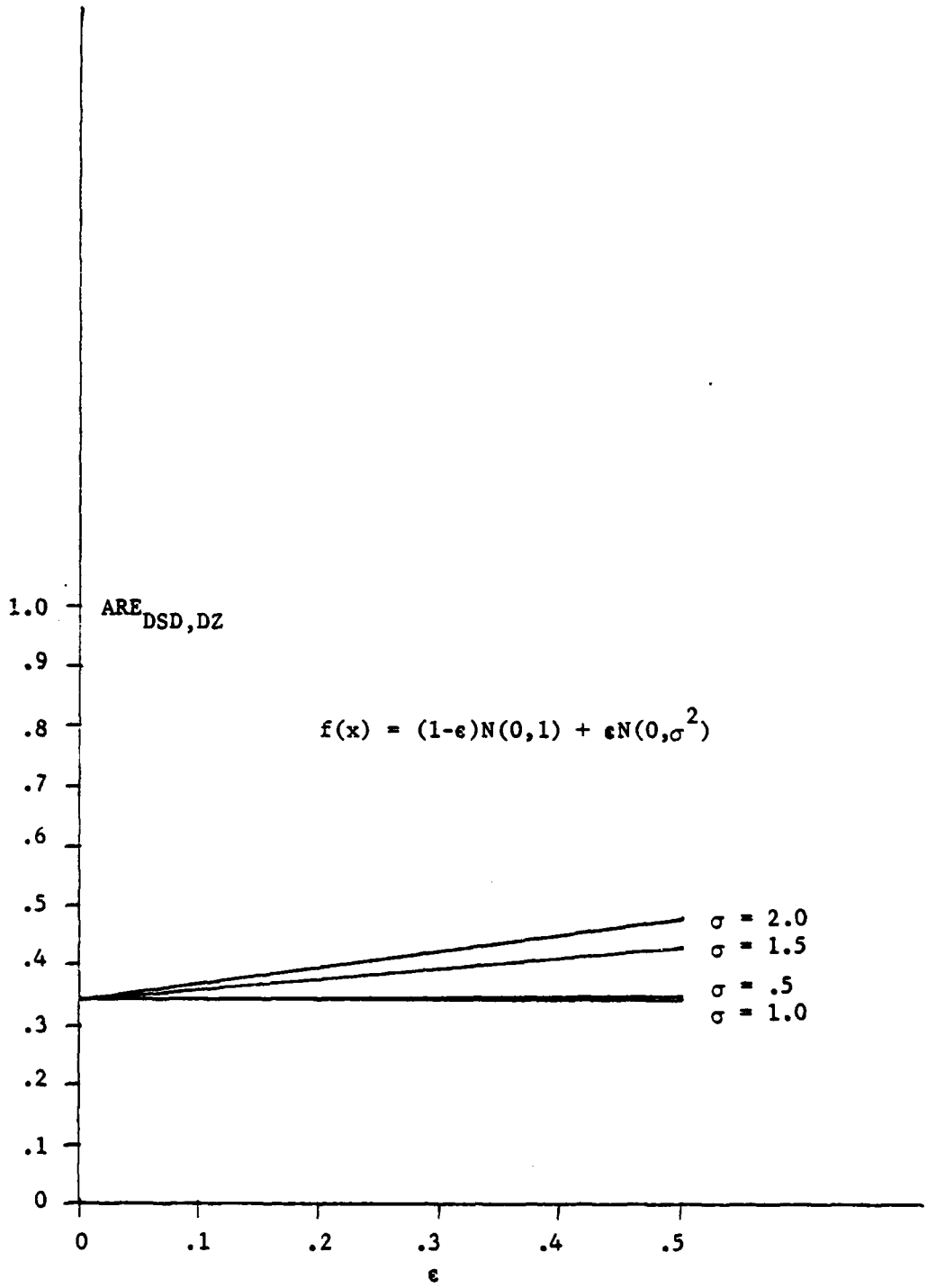


Fig. 18. Dithered sign detector compared with dead-zone quantizer.

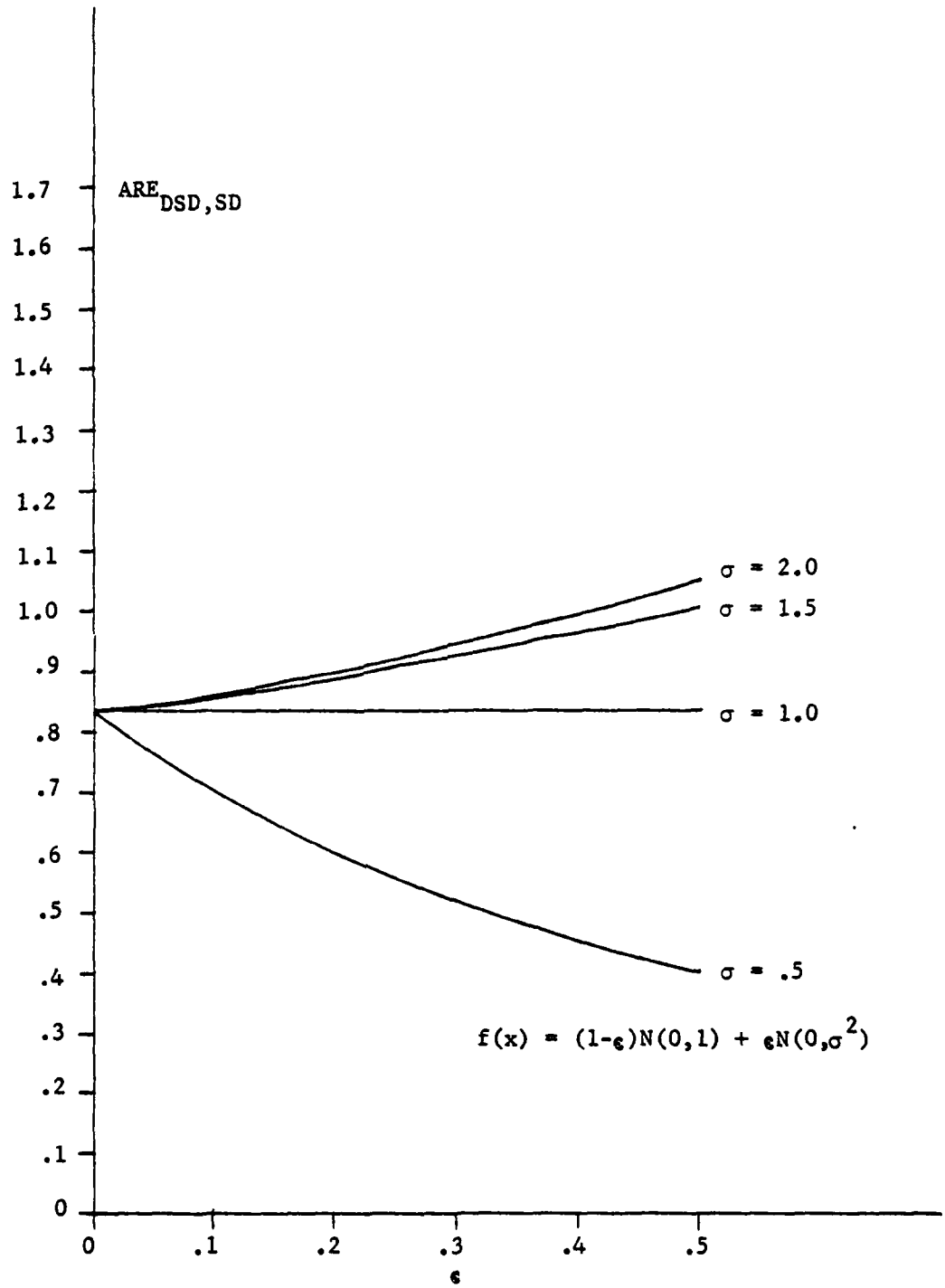


Fig. 19. Dithered sign detector compared with sign detector.

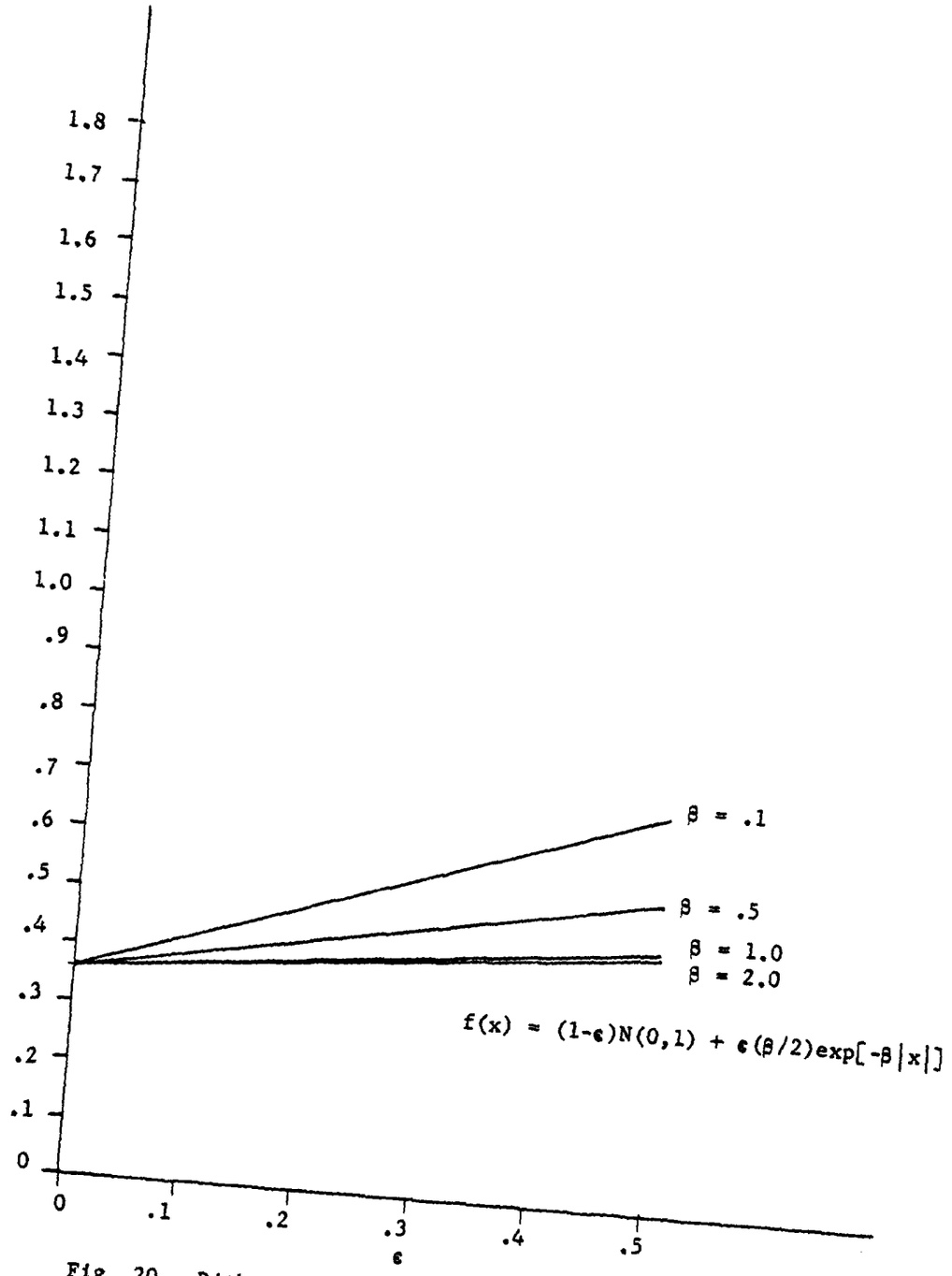


Fig. 20. Dithered sign detector compared with dead-zone.

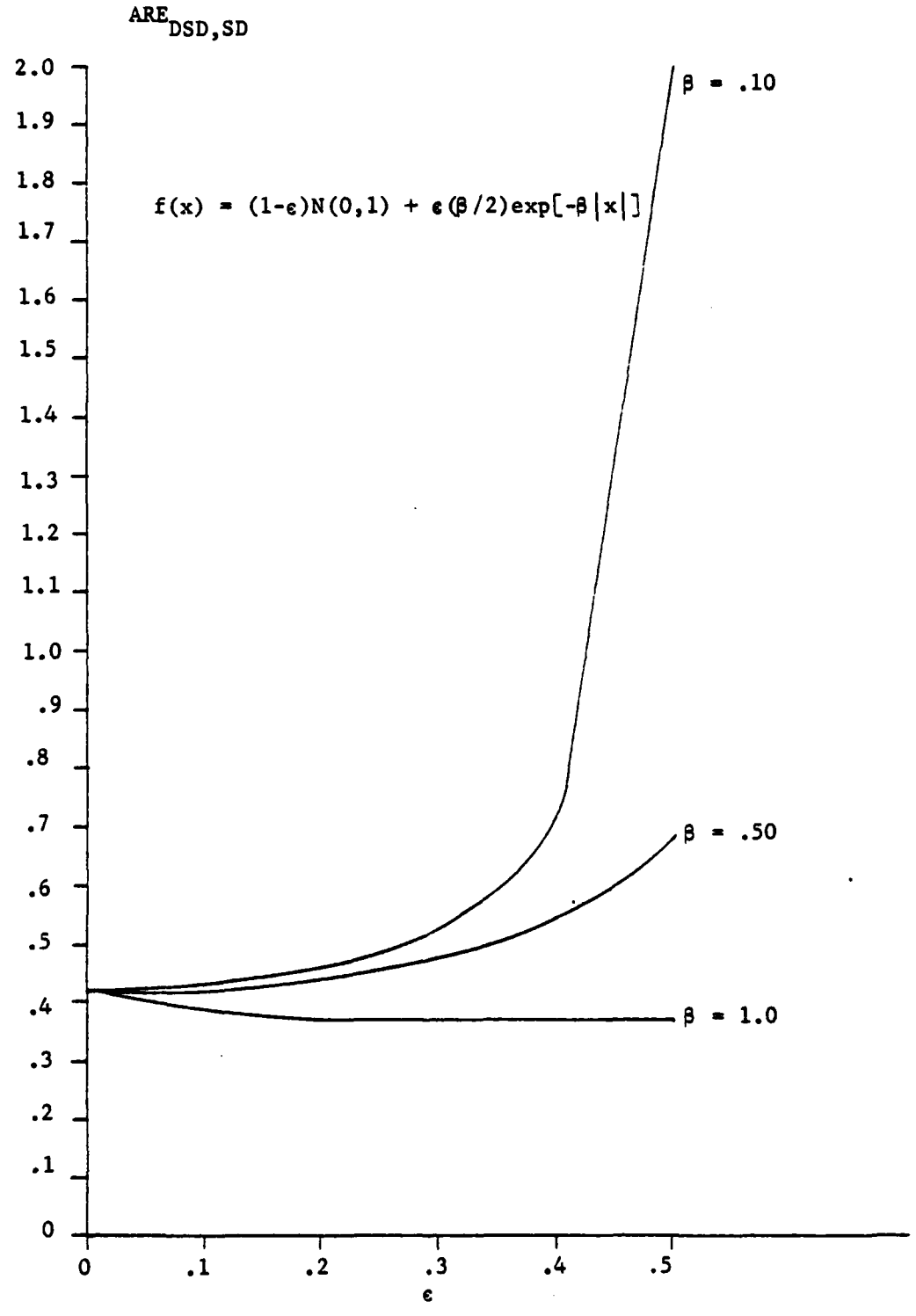


Fig. 21. Dithered sign detector compared with sign detector.

It appears from these results that a dithered dead-zone quantizer may be useful in comparison with an ordinary dead-zone quantizer for severe contamination of Gaussian noise with exponentially peaked noise densities. This appears to be restrictive in the light of the reason for proposing the dithered system - namely to improve robustness. If robustness indicates relative insensitivity to changes in the noise it seems wasteful to use a system of increased complexity to gain only moderate performance improvement over a restricted range of noise contamination.

4. THE 2m-LEVEL QUANTIZER AS AN APPROXIMATION TO THE LIMITER CORRELATOR

Our last investigation of a quantizer-detection scheme is possibly the most general. Here we compare a quantizer with 2m levels (m is an integer) with the limiter-correlator as proposed by Martin and Schwartz [2]. The motivation here is that the quantizer (which is uniform and symmetrical) appears as a discrete approximation to the limiter correlator and represents the approximation which would be achieved by a digital computer implementation of the detector.

First we will reiterate the development of the limiter-correlator as in Martin and Schwartz. The problem is one of deciding between a hypothesis and alternative H_0 and H_1 as described in the Introduction. The noise density $f(\cdot)$ is assumed to belong to the family of densities

$$\mathcal{G} = \{f(x) \mid f(x) = (1-\epsilon)\varphi(x) + \epsilon h(x) : h(\cdot) \in \mathcal{K}, \quad 0 \leq \epsilon < 1\} \tag{4.1}$$

where $\varphi(x)$ denotes the unit variance Gaussian density with zero mean, $\mathcal{K}$ is the class of all symmetric density functions satisfying the following regularity condition:

Regularity Condition: Let $I(\theta) = \int \ell(x)h(x-\theta)dx$ where $\ell(\cdot)$ is a bounded function and $h(\cdot) \in \mathcal{K}$. Then $I(\theta)$ has one continuous derivative given by

$$I'(\theta) = \int \ell(x)\,(\partial h(x-\theta)/\partial\theta)dx$$

Since $\varphi(x)$ satisfies this condition as well as $h(x)$, then $\mathcal{G}$ also satisfies the regularity condition.

Note that the other mixture models used throughout this thesis (with $h(x) = N(0,\sigma^2)$ and $h(x) = (\beta/2)\exp[-\beta|x|]$) also meet this condition.

Let $\mathcal{A}$ denote the set of all randomized test functions $\phi = \phi(X)$ where $X = (x_1,\ldots,x_N)$ is the set of N observations. That is, $\phi(X)$ represents the probability of choosing H_1 as true given that the observation is X.

The power function $\beta_\phi(\theta|f)$ of $\phi \in \mathcal{A}$ is the probability of choosing H_1 given that H_1 is true using the randomized test function ϕ. This is given by

$$\beta_\phi(\theta|f) = E_\theta[\phi(X)|f] \qquad f \in \mathcal{F} \tag{4.2}$$

The false-alarm probability is

$$\alpha = \beta_\phi(0|f) \tag{4.3}$$

The problem is to find a locally most powerful test $\tilde{\phi}$ and a least-favorable density $\tilde{f}$ such that the following conditions for the power are true:

$$\beta_{\tilde{\phi}}(0|f) \leq \beta_{\tilde{\phi}}(0|\tilde{f}) \tag{4.4}$$

$$\beta'_{\tilde{\phi}}(0|f) \geq \beta'_{\tilde{\phi}}(0|\tilde{f}) \tag{4.5}$$

where $$\beta'_{\tilde{\phi}}(0|\cdot) \triangleq (d/d\theta)\beta_{\tilde{\phi}}(\theta|\cdot)\Big|_{\theta=0}$$

If $\tilde{\phi}$ is indeed the locally most powerful test for $\tilde{f}$, then $\tilde{\phi}$ and $\tilde{f}$ form a saddle-point solution in terms of local power with constrained false alarm probability and (4.6) and (4.7) below are satisfied.

$$\inf_{f \in \mathcal{F}} \beta'_{\tilde{\phi}}(0|\tilde{f}) = \beta'_{\tilde{\phi}}(0|\tilde{f}) = \sup_{\phi \in \mathcal{A}} \beta'_{\phi}(0|\tilde{f}) \tag{4.6}$$

$$\alpha = \beta_{\tilde{\phi}}(0|f) \leq \beta_{\tilde{\phi}}(0|\tilde{f}) = \tilde{\alpha} \tag{4.7}$$

Now $\tilde{f}$ and $\tilde{\phi}$ must be specified for each ϵ such that (4.6) and (4.7) hold asymptotically over a restricted range of $\tilde{\alpha}$ bounded away from zero and dependent on ϵ.

The least-favorable density $\tilde{f}$ is given by

$$\tilde{f}(x) = (1-\epsilon)^{-\frac{1}{2}}\exp[-g(x)]$$

$$g(x) = \begin{cases} x^2/2 & |x| < K \\ K|x| - K^2/2 & |x| \geq K \end{cases} \tag{4.8}$$

or, since $\tilde{f}(x) = (1-\epsilon)\varphi(x) + \epsilon\tilde{h}(x)$

$$\tilde{h}(x) = \begin{cases} 0 & \\ (2\pi\epsilon^2)^{-\frac{1}{2}}(1-\epsilon)[\exp(-K|x|+K^2/2)-\exp(-x^2/2)] & |x| \geq K \end{cases} \tag{4.9}$$

In addition, if $\tilde{h}(x)$ is to be a valid density function (i.e. if $\int_{-\infty}^{\infty} \tilde{h}(x)dx = 1$), then K satisfies

$$\int_{-K}^{K} \varphi(x)dx + 2(\varphi(K)/K) = (1-\epsilon)^{-1} \tag{4.10}$$

Now for a noise density given by $\tilde{f}$, the density for an observation, x_i, given H_1 is true, is

$$\tilde{f}_\theta(x_i) = \tilde{f}(x_i - \theta s_i) \tag{4.11}$$

According to the generalized Neyman-Pearson Lemma [15] then, the locally most powerful test $\tilde{\phi}$ of size $\tilde{\alpha}$ is given by

$$\tilde{\phi}_N(X) = \begin{cases} 1 & T_N(X) > c \\ a & T_N(X) = c \\ 0 & T_N(X) < c \end{cases} \tag{4.12}$$

where $0 \leq a \leq 1$

and $$T_N(X) = \sum_{i=1}^{N} s_i \ell(x_i; -K, K) \tag{4.13}$$

Here, $\ell(y;A,B)$ indicates a soft limiter characteristic given by (4.14) below

$$\ell(y;A,B) = \begin{cases} B & y \geq b \\ y & A < y < B \\ A & y \leq A \end{cases} \tag{4.14}$$

This characteristic can be derived from the derivative of the probability ratio test using $\tilde{f}(x)$ above, i.e. the expression of (4.12) is equivalent to

$$\frac{\partial}{\partial\theta} \prod_{i=1}^{n} \frac{\tilde{f}(x_i-\theta s_i)}{\tilde{f}(x_i)}\Bigg|_{\theta=0} \gtrless \tau \tag{4.15}$$

where τ is a threshold chosen to achieve a desired probability of false alarm $\tilde{\alpha}$. Constant terms are includes as a new threshold c.

The resulting detector structure has been dubbed the limiter-correlator and is illustrated in Fig. 22. Note that the limiter is symmetrical and that large ϵ increases the limiting action (i.e., reduces K).

To conclude the description of the correlator-limiter we include Theorem 2 of [2].

Define

$$\gamma_m = (A + \epsilon K^2)/A \tag{4.16}$$

where

$$A = (1-\epsilon)\int_{-K}^{K} x^2\varphi(x)dx + 2K^2\Phi(-K) \tag{4.17}$$

Since K is a function of ϵ, γ_m is also, therefore $\gamma_m = \gamma_m(\epsilon)$ and α is then also a function of ϵ given in terms of γ_m:

$$\alpha(\epsilon) = 1 - \Phi([2\log\gamma_m/\gamma_m^2 - 1)]^{\frac{1}{2}}) \tag{4.18}$$

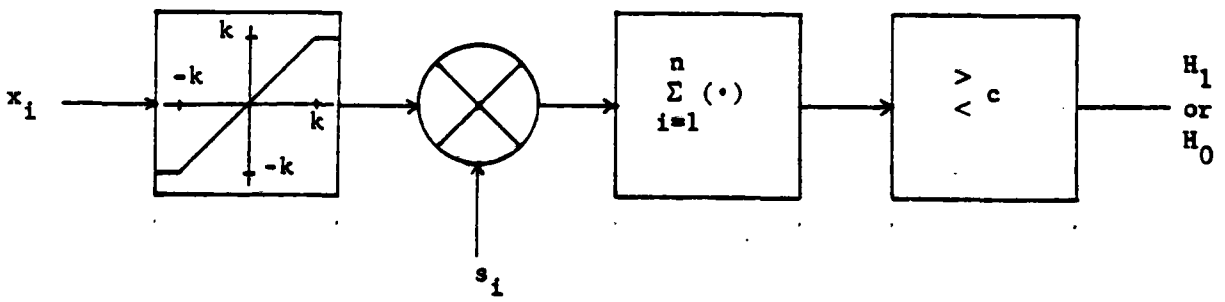


Fig. 22. Limiter-correlator detector.

Theorem : Let $\tilde{f}$ be the least favorable density (29) and $\tilde{\phi}_N$ the corresponding locally most powerful test given by (33) and (34). If $\tilde{\alpha} \geq \alpha(\epsilon)$, $0 < \epsilon < 1$, then asymptotically

$$\inf_{f \in \mathcal{J}} \beta'_{\tilde{\phi}_N}(0|f) = \beta'_{\tilde{\phi}_N}(0|\tilde{f}) = \sup_{\tilde{\phi} \in \mathcal{S}} \beta'_{\tilde{\phi}_N}(0|f)$$

and

$$\beta_{\tilde{\phi}_N}(0|f) \leq \beta_{\tilde{\phi}_N}(0|\tilde{f}) = \tilde{\alpha}$$

A comparison of the limiter characteristic of Fig. 22 and the typical quantizer of Fig. 2 reveals the similarities between the two. If $q_m = K$, then the quantizer may be considered a discrete approximation to the limiter characteristic. As the number of quantizer steps grows, i.e. as m is increased, and as long as $q_m = K$, the steps become finer and the quantizer becomes a closer approximation to the limiter. From experience with the quantization of analog signals in the ordinary A/D, D/A case of signal processing, one expects more steps to create an improvement in performance, upper bounded by the performance of the continuous (or piecewise-continuous) system to which the quantizer is an approximation.

In standard data quantization, the comparison between a system function and its quantized version is mean-squared-error [7], absolute-mean-error [11], or some more general error criterion [12,13]. These methods do not always lead to the best solution when one considers a quantizer in a binary decision system, however, and some other criteria may give better results, e.g. distance measures [14].

To reiterate then, the best detection quantizer may not be merely the least-mean-square-error approximation to the analog optimum systems. In

light of this fact, the following questions can be asked concerning the performance of the 2m-level uniform quantizer as an approximation to the limiter-correlator:

1. What is the ARE of the quantizer compared to the limiter ($ARE_{Q,LC}$)?
2. How does randomization affect the $ARE_{Q,LC}$?
3. How does dithering affect the $ARE_{Q,LC}$?

Implicit in theses queries is the effect of m on the $ARE_{Q,LC}$.

Why, if this quantizer may be suboptimal as a detector are we worried about its performance compared to the analog system? First of all, signal processing of any complexity is quite likely to be carried out digitally, implying a quantization of data. Secondly this quantization is likely initially to be of the uniform, 2m-level type since this is the nature of most commercially available A/D converters. Thus the need for a nonuniform quantizer might overly complicate a system for little gain in terms of detector performance. Thus the answers to the three questions above could well indicate the economy of constructing a more complicated quantizer.

The efficacy of the limiter-correlator is given by

$$\eta_{LC} = \frac{\left[\int_{-K}^{K} f(n)dn\right]^2}{2K^2 \int_{-\infty}^{-K} f(n)dn + \int_{-K}^{K} n^2 f(n)dn} \tag{4.19}$$

and that of the 2m-level uniform quantizer by (see Appendix A):

$$\eta_Q = \frac{2\left[\sum_{k=1}^{m-1} (2k-1)\{f((k-1)t) - f(kt)\} + (2m-1)f((m-1)t)\right]^2}{\sum_{k=1}^{m-1} (2k-1)^2\{F(kt) - F((k-1)t)\} + (2m-1)^2\{1-F((k-1)t)\}} \tag{4.20}$$

If $q_m = K$, then $t = K/(m-1/2)$ to maintain the similarity between the two detectors.

The results here are multidimensional. ARE's are calculated and plotted for both the Gaussian contaminated and the exponential contaminated noise with m = 3,4 and 5. In addition this is carried out for the least-favorable noise density. All of these results use ϵ and K as parameters to form a set of four curves for each m in each noise case. Plotting the ARE as a function of ϵ was not feasible due to the complicated relationship between ϵ and K as evidenced by (4.10). The four paired values of ϵ and K are listed in Table 2 below.

Table 2

ϵ	.01	.02	.05	.10
K	1.95	1.72	1.40	1.14

Figures 23 through 25 show the quantizer performance in the Gaussian contaminated noise. Surprisingly $ARE_{Q,LC} > 1$ for all cases considered. Two trends are clearly evident: the quantizer performance improves over the limiter-correlator for more severe mixtures (ϵ increasing), and as m increases, the quantizer perfcrmance decreases. Also, at first glance, one would expect the performance for all mixtures to be equal at $\sigma = 1$, however, it must be remembered that K changes with ϵ so that even though m remains the same, the quantizers are not equivalent (nor are the limiters) for different ϵ. The improvement obtained here is significant and is quite probably worth the implementation of a quantizer.

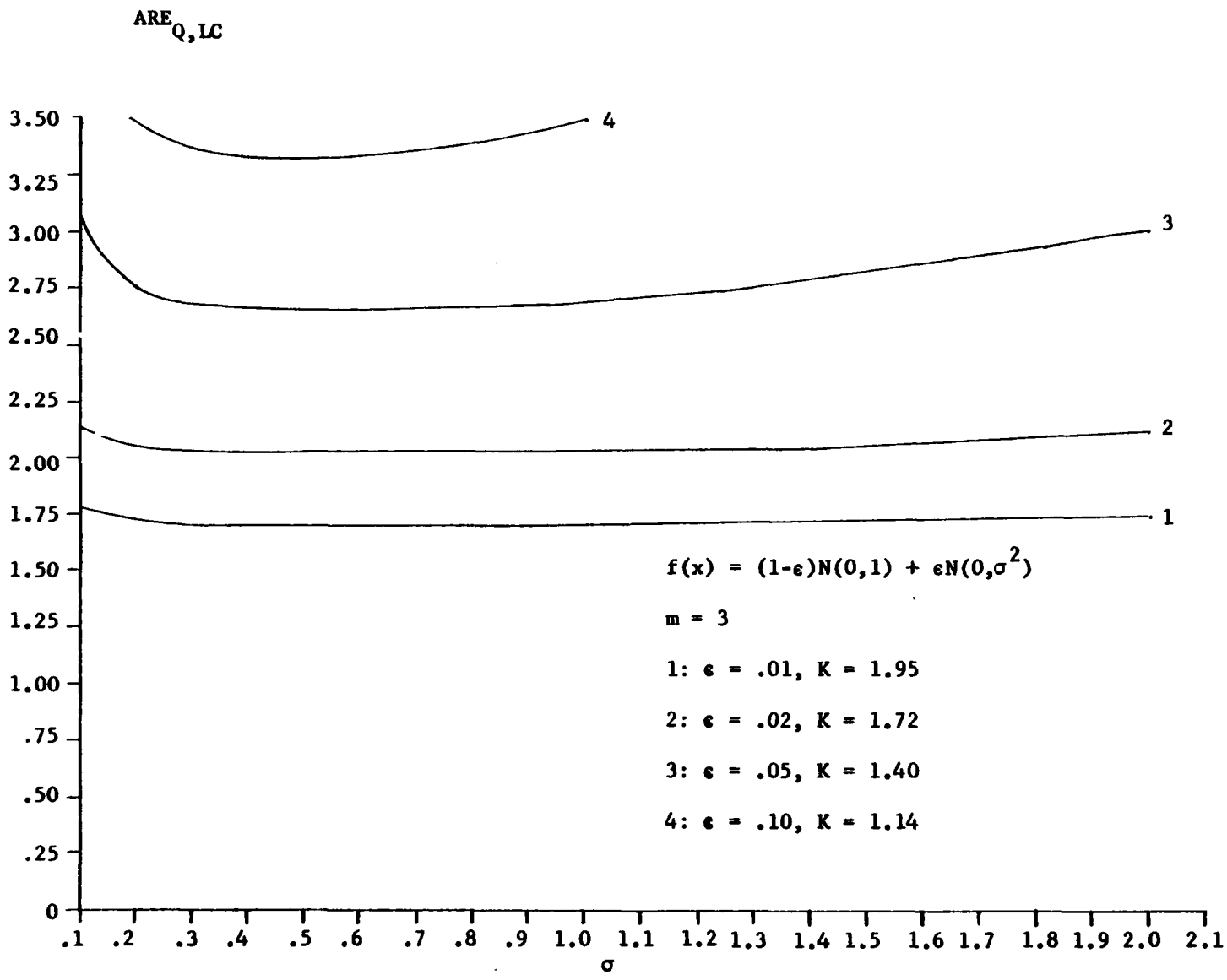


Fig. 23. Comparison of quantizer with limiter-correlator.

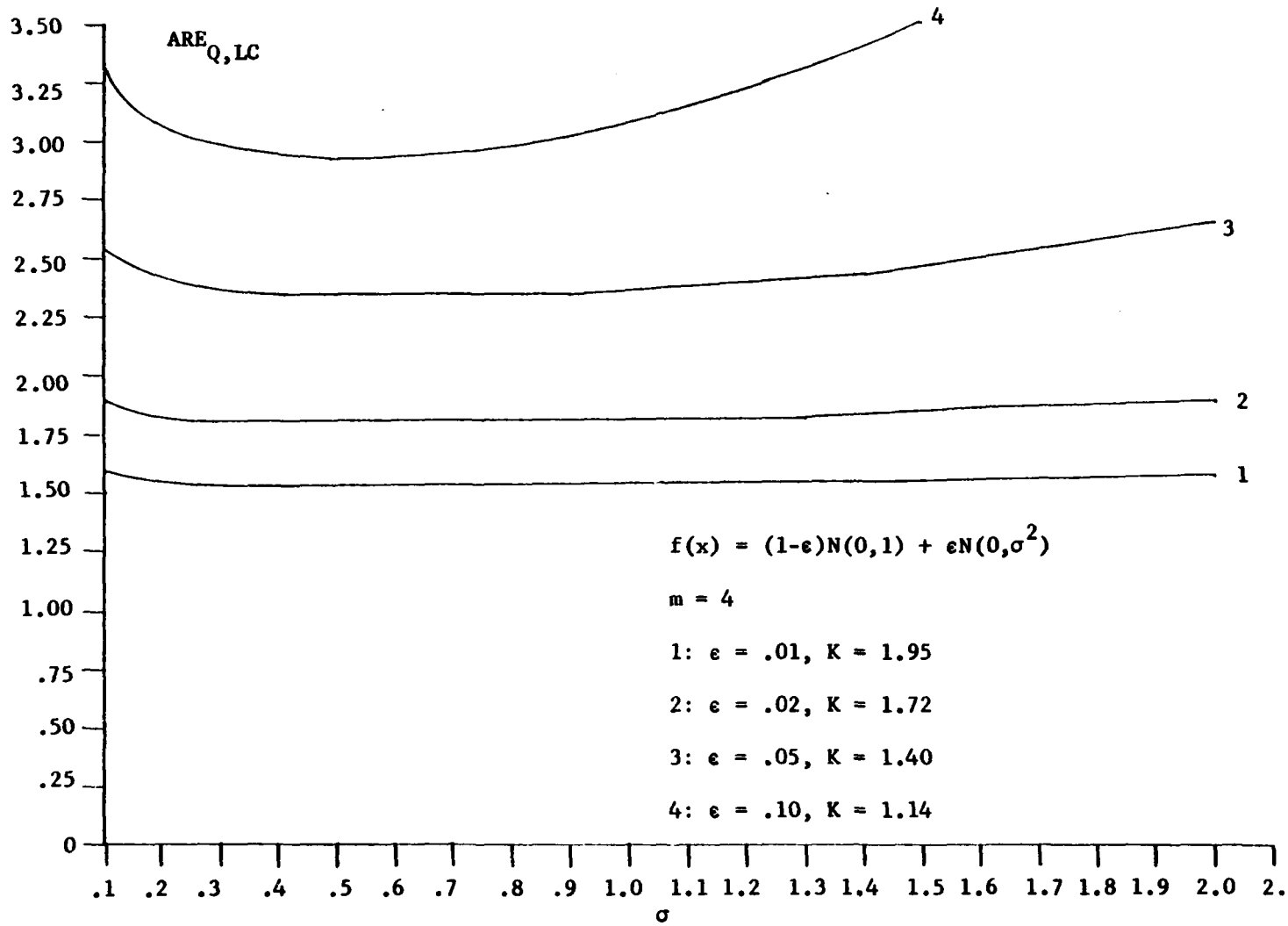


Fig. 24. Comparison of quantizer with limiter-correlator.

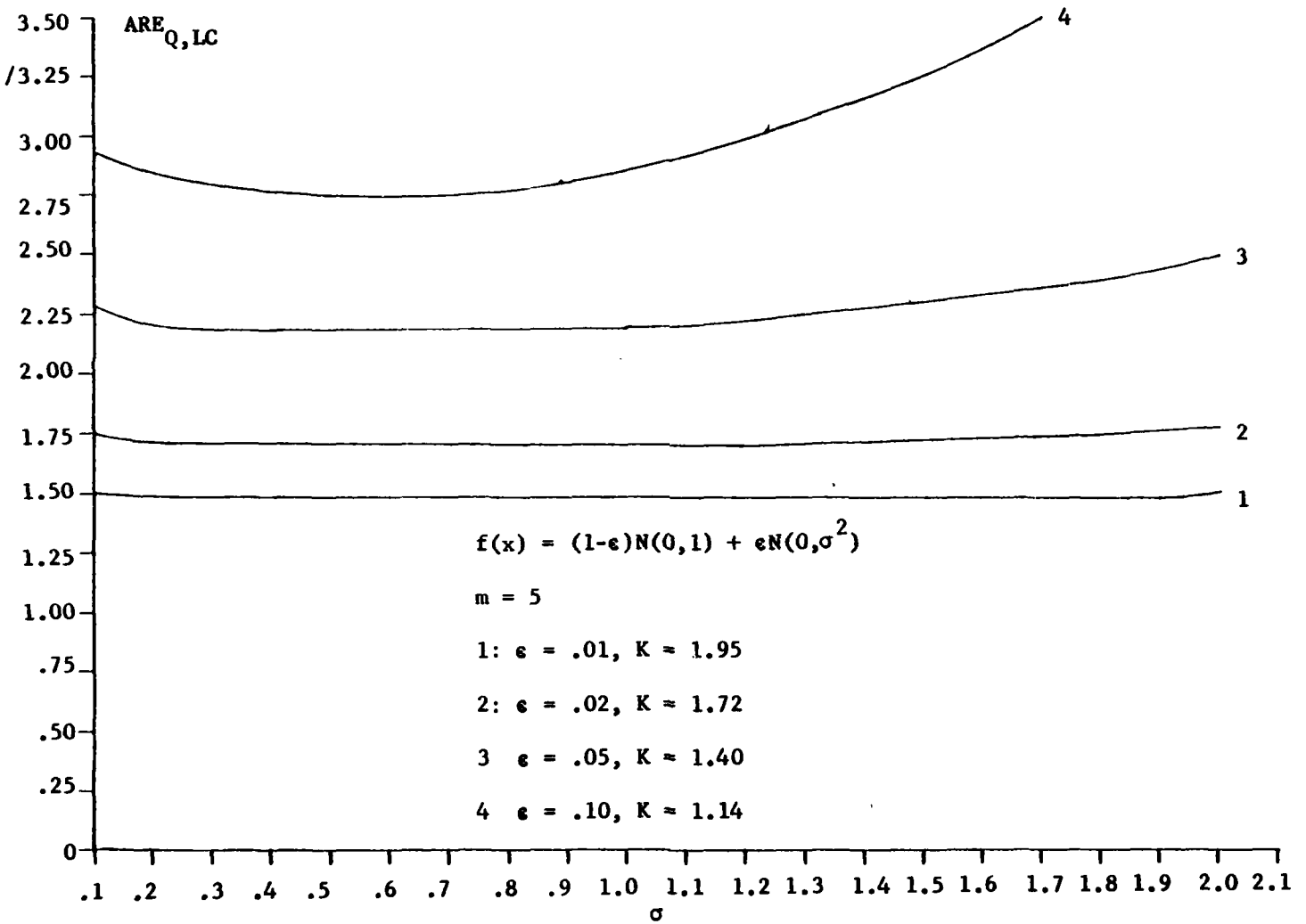


Fig. 25. Comparison of quantizer with limiter-correlator.

The next question is how does randomization affect the ARE? In this case the intervals $[t_{k-1},t_k)$ are treated as random variables all with identical uniform densities. To be more precise, we define u to be the length of the interval $[t_{k-1},t_k)$, a constant for all values of k: $u = t_k - t_{k-1}$. We then make u a random variable with a uniform density function over (0,c). The effect is a quantizer which collapses and expands as u takes on smaller and larger values respectively. Note the level values q_k are not changed. If u = 0, we have a sign detector with output levels $\pm q_m$. If u = c, the output levels consist of the set $\{-q_m,\ldots,-q_1,0,q_1,\ldots,q_m\}$ with breakpoints $t_k = kc$.

For these results, $c = 2t_1$ where t_1 is the first breakpoint of the corresponding non-randomized quantizer, i.e. $c = 2K/(m-1/2)$.

Given the above density function for u we have an expression for the efficacy of the randomized 2m-level quantizer given by (see Appendix B).

$$\eta_{RQ} = \frac{\frac{2}{c}\left[\sum_{k=1}^{m-1}(2k-1)\int_0^c \{f((k-1)x) - f(kx)\}dx + (2m-1)\int_0^c f((m-1)x)dx\right]^2}{\sum_{k=1}^{m-1}(2k-1)^2\int_0^c \{F(kx) - F((k-1)x)\}dx + (2m-1)^2\int_0^c \{1-F((m-1)x)\}dx} \qquad (4.21)$$

Figures 26, 27, and 28 illustrate the performance for m = 3,4 and 5 respectively. Obviously randomizing the interval is not a technique useful for improving performance in this case. Once again we notice a tendency for the $ARE_{RQ,LC}$ to improve for small σ indicating some improvement by randomizing for peaked contamination noise. However for this to compete with the nonrandomized quantizer, σ must be very small.

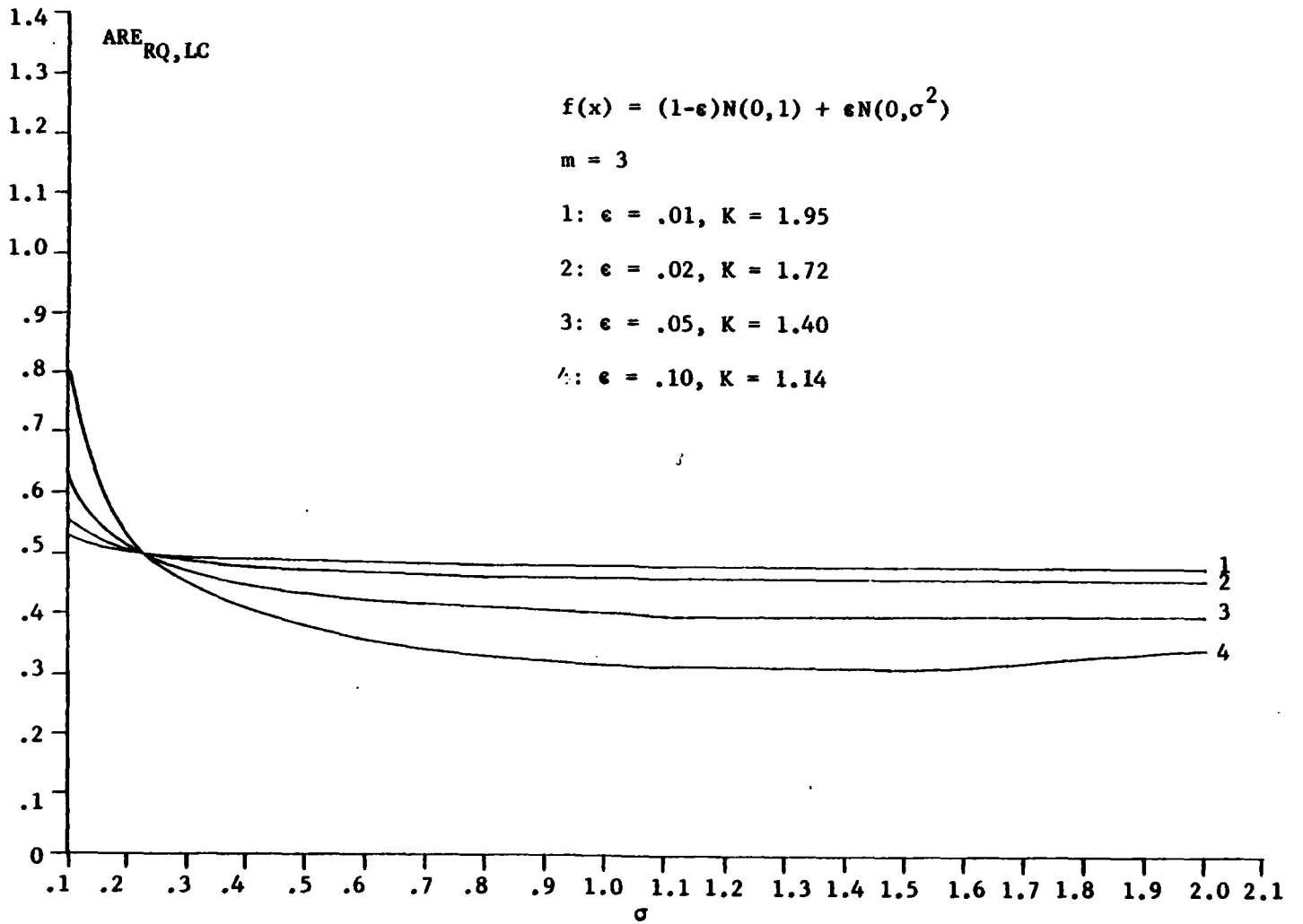


Fig. 26. Comparison of randomized quantizer with limiter-correlator.

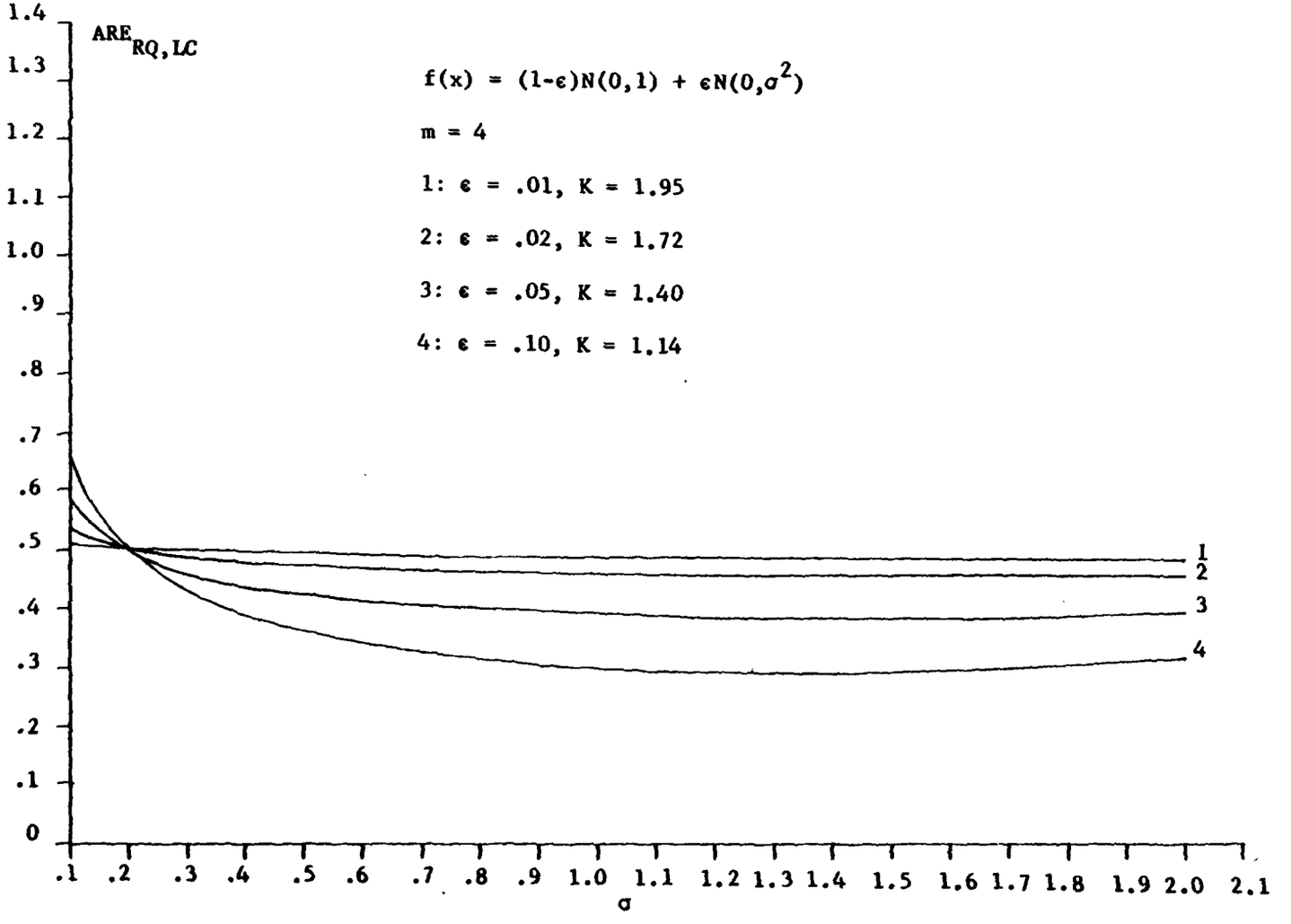


Fig. 27. Comparison of randomized quantizer with limiter-correlator.

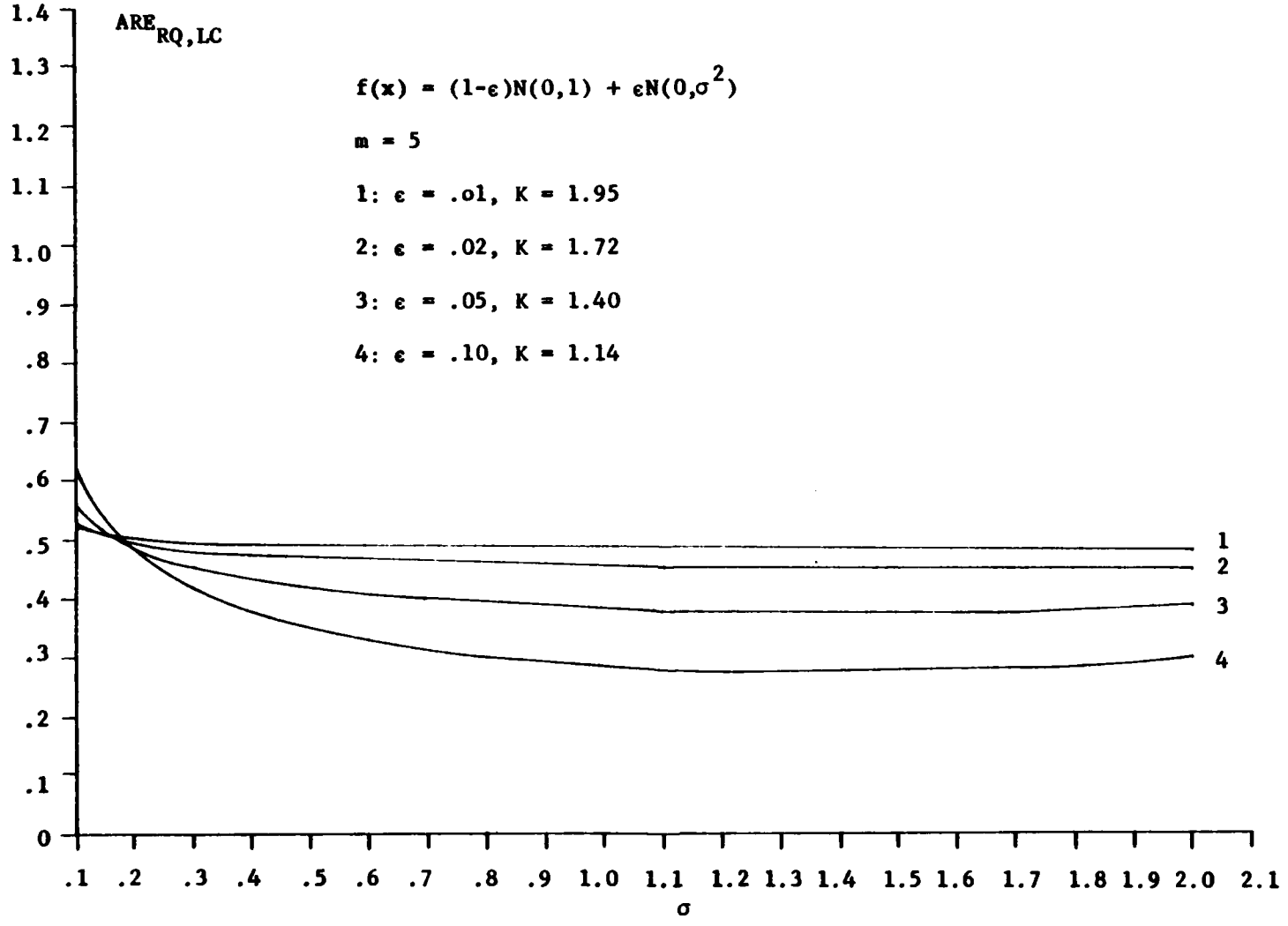


Fig. 28. Comparison of randomized quantizer with limiter-correlator.

Figures 29, 30, and 31 are the performance curves of the non-randomized quantizer for m = 3,4,5 in exponential contaminated noise. Figures 32, 33, and 34 are the corresponding curves for the randomized quantizer. The observations here are much the same as for the Gaussian contamination case - the quantizer performance is markedly better than the limiter and randomization ruins this performance. In this case, randomization produces uniformly worse results with no promise of an increase in $ARE_{RQ,LC}$ for any β or ϵ.

Randomization produces such terrible results that one would be led to expect a similar degradation with dithering.

In calculating the efficacy of the dithered system, the variance given H_0 must be found. This is given by

$$\mathrm{Var}[T_N(X)|H_0] = \sum_{i=1}^{n} s_i^2\{\mathrm{Var}[Q(n_i+d_i)|H_0] + \frac{t^2}{12} - 2E[d_i \cdot Q(n_i+d_i)|H_0]\} \quad (4.22)$$

where n_i = noise samples

d_i = dither signal samples

$Q(x)$ = output of quantizer with input x

(Note: Appendix C contains a general solution for this term).

Here d_i has a uniform density function over (-t/2, t/2).

The last term in (4.22) is the correlation between d_i and the quantizer output. The effect of d_i is to cause $Q(n_i + d_i)$ to change by only a single step value up or down from what $Q(n_i)$ would be, if d_i causes a change at all. Thus, although d_i and $Q(n_i + d_i)$ are not independent or uncorrelated, we can see that the correlation is small and positive. Since the calculation of this term becomes very complicated we will take it to be zero. This

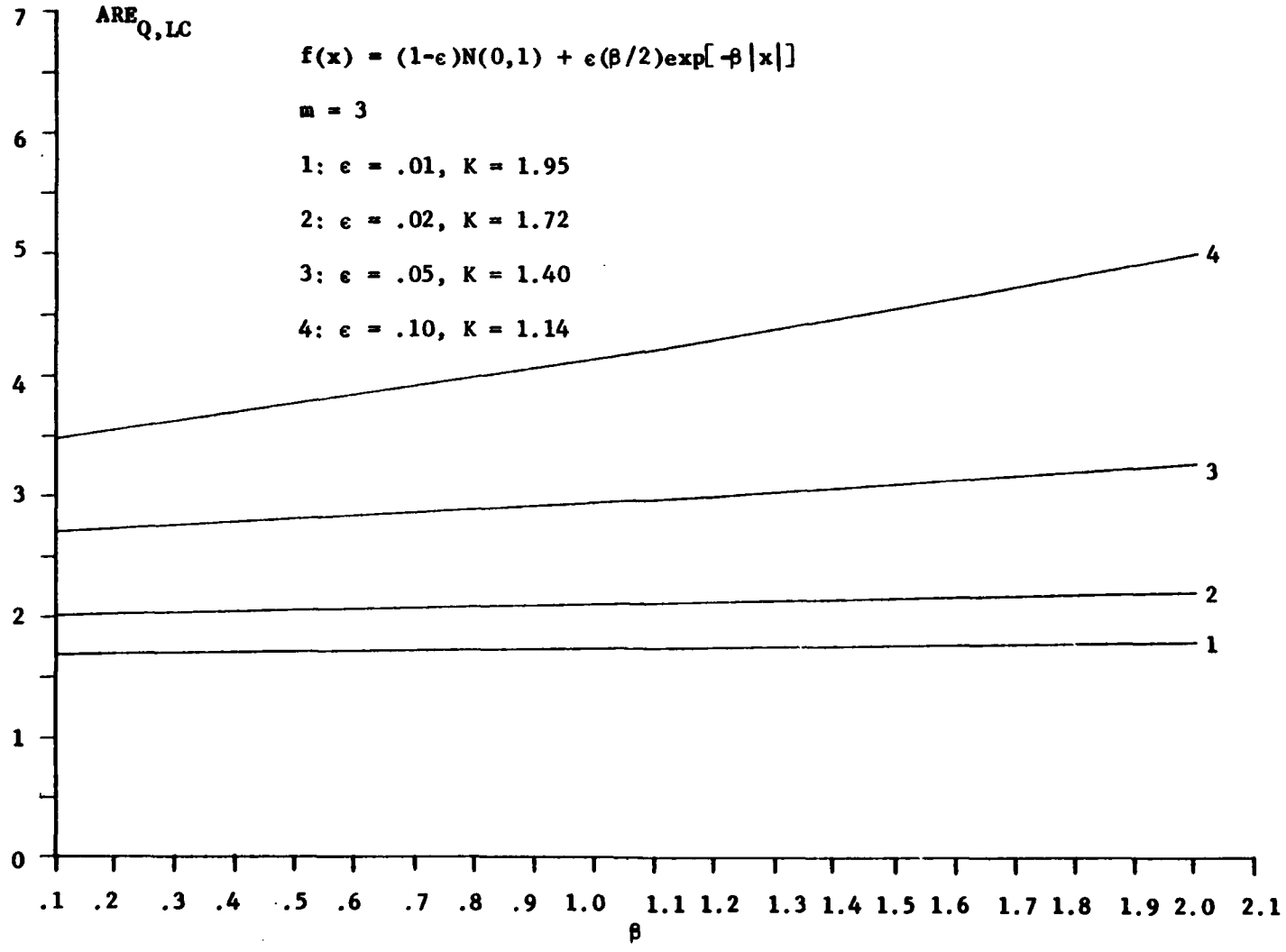


Fig. 29. Comparison of quantizer with limiter-correlator.

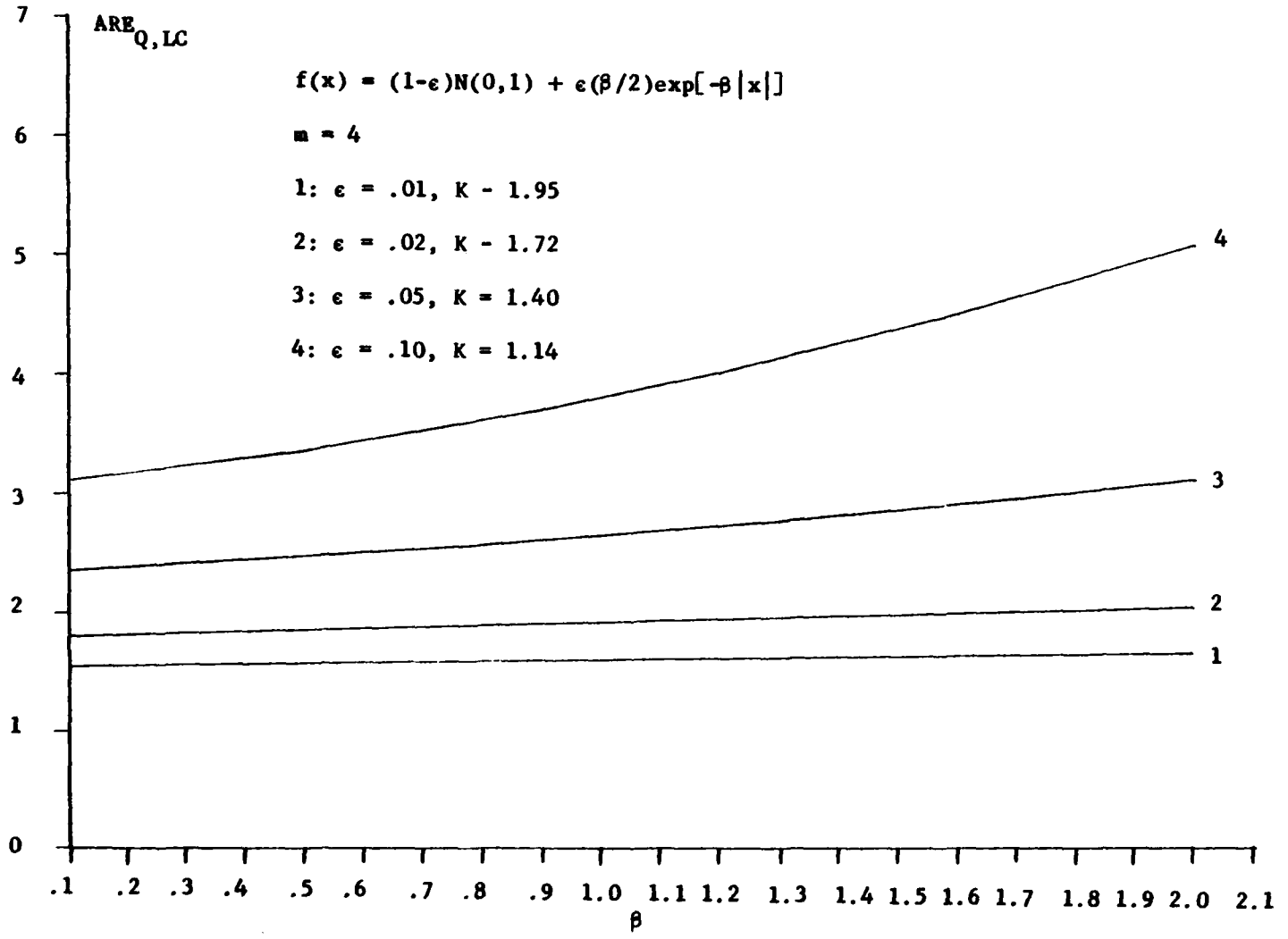


Fig. 30. Comparison of quantizer with limiter-correlator.

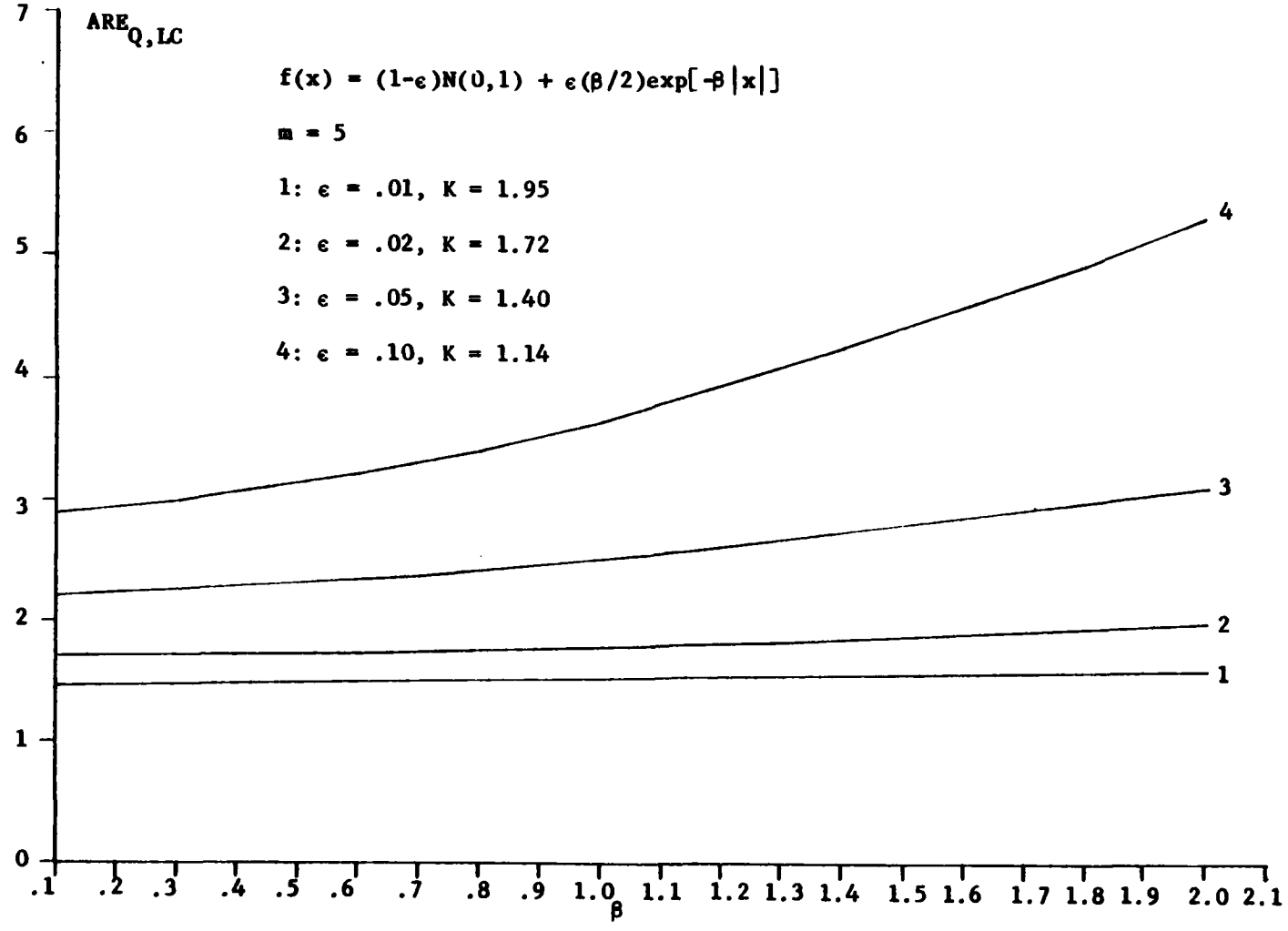


Fig. 31. Comparison of quantizer with limiter-correlator.

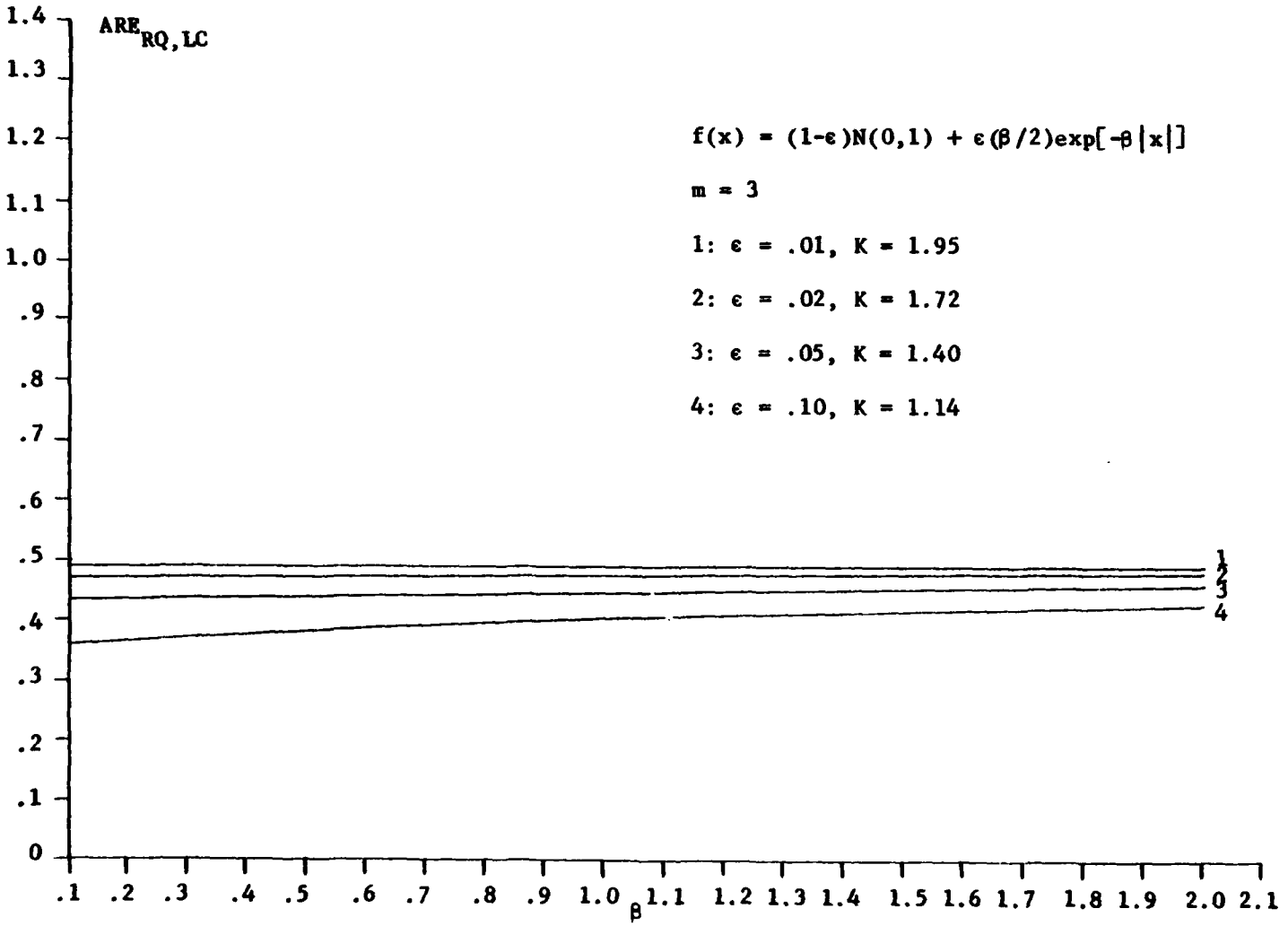


Fig. 32. Comparison of randomized quantizer with limiter-correlator.

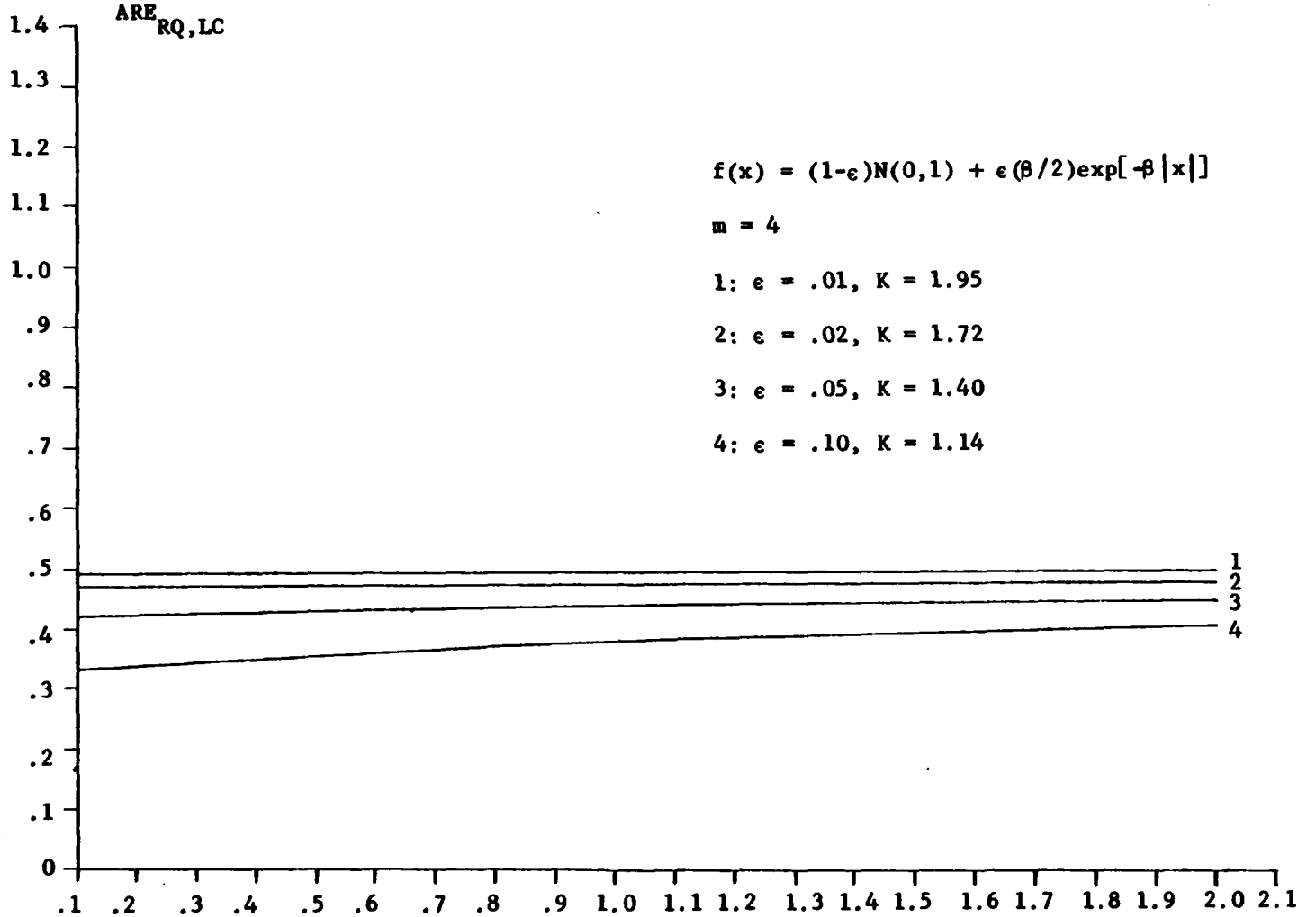


Fig. 33. Comparison of randomized quantizer with limiter-correlator.

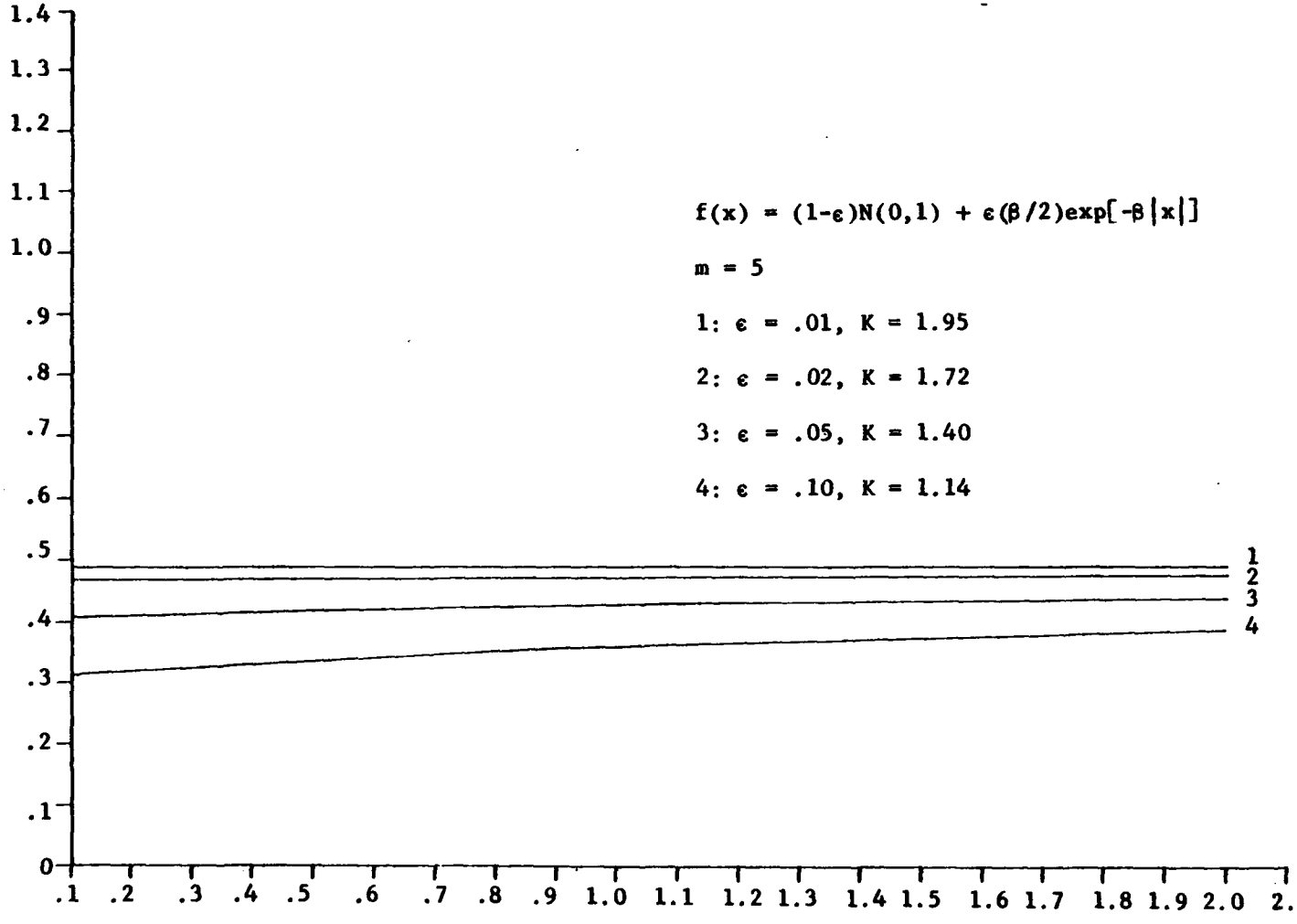


Fig. 34. Comparison of randomized quantizer with limiter-correlator.

establishes a lower bound on the efficacy of the dithered quantizer since inclusion of the term can only increase the value of η_{DQ}.

Under this assumption the efficacy of the dithered quantizer is given by

$$\eta_{DQ} \cong \frac{2\left[\sum_{k=1}^{m}(2k-1)\{2F((k-\frac{1}{2})t) - F((k-\frac{3}{2})t) - F((k+\frac{1}{2})t)\}\right]^2}{t\sum_{k=1}^{m}(2k-1)^2\left[\int_{(k-1)t}^{kt}\{F(z_i+\frac{t}{2}) - F(z_i-\frac{t}{2})\}dz_i + \frac{t^2}{12q_1^2}\right]} \tag{4.23}$$

where $z_i = n_i + d_i$. Note that once again, dithering brings the output level q into play.

Figure 35 shows the best $ARE_{DQ,LC}$ curve obtained with the exponential contaminated noise. It appears that no advantage is gained by dithering here. As ϵ increases, so does the ARE, but again the values at which any performance gain appears are too large for a valid noise model.

As a final check of the validity of the results for the 2m-level quantizer we examine the $ARE_{Q,LC}$ when the noise density is given by (4.8), the least favorable density. Since the limiter-correlator is the optimum detector for this case, we expect $ARE_{Q,LC} < 1$ for all ϵ. The values obtained are summarized in Table 3.

Table 3

$ARE_{Q,LC}$

ϵ	m = 3	m = 4	m = 5
.01	.952	.975	.985
.02	.962	.980	.988
.05	.975	.987	.992
.10	.983	.991	.995

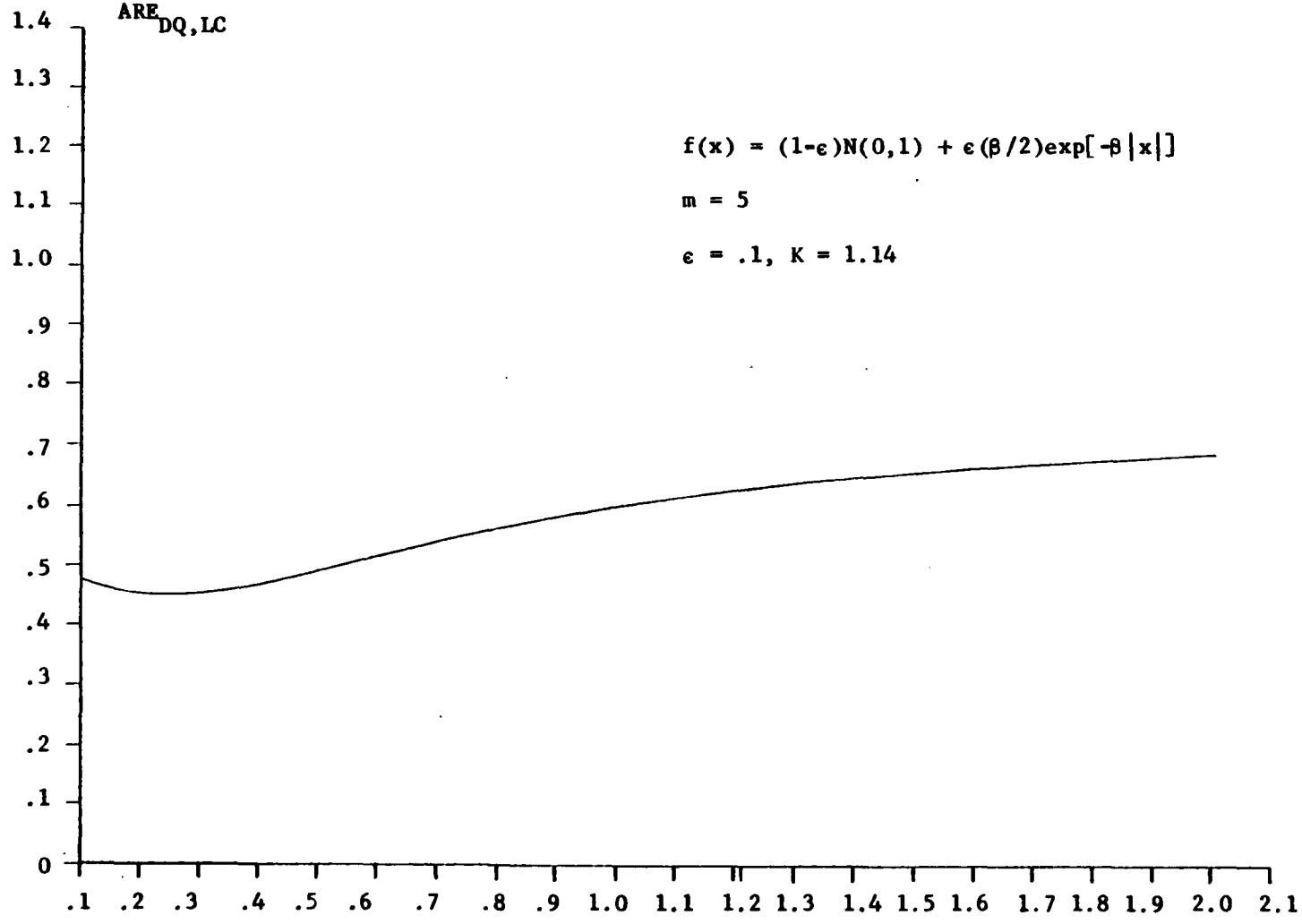


Fig. 35. Comparison of dithered quantizer with limiter-correlator.

This reinforces the optimality of the limiter correlator for the least-favorable density as well as verifying the validity of previous results since the same general equations for efficacy were used throughout this section. Also we see that, as the number of steps increases, the performance of the quantizer approaches that of the limiter-correlator. The performance also increases with increasing ϵ which is consistent with earlier findings. Thus it appears that in cases where the limiter-correlator is optimal, our intuition is correct in assuming that increasing the number of steps will improve the effectiveness of a quantizer used to replace the limiter.

The answers to the three questions posed earlier can now be stated. A quantizer may be used to some advantage in place of the limiter. Problems arise however in trying to design the quantizer. If the noise is close to the least favorable noise in some sense, we would do well to provide as many steps as possible to approach the performance of the limiter as closely as possible. If on the other hand, the noise is quite different (e.g. Gaussian or exponential contamination) then the quantizer should have relatively few steps and we can expect greatly improved performance. In either case, of course, there is also the problem of choosing an adequate mixture model.

As far as randomization and dithering go, it appears that both lead to either uniformly poor results or results which are beneficial under such narrow conditions that the question of robustness is moot.

The best that can be done then, is to choose a noise model (for the mixture model) and to determine for each choice whether or not the quantizer replacement for the limiter is practical. There does not appear to be any rule of thumb concerning this choice.

5. SUMMARY

We have examined quantizers in three different forms: the four-level symmetrical uniform quantizer, the three-level or dead-zone quantizer and the more general 2m-level symmetrical uniform quantizer. For all three types the performance of the quantizer as a signal processor in a binary hypothesis testing scheme was compared with several more widely used detector nonlinearities known for their performance stability with respect to slight noise changes.

The object of these calculations was to observe how, if at all, the quantizers performed in terms of this stability, called robustness. A contaminated noise mixture and the asymptotic relative efficiency (ARE) were used to measure this robustness relative to that of the aforementioned well known nonlinearities (the sign detector and the limiter-correlator).

In addition, two techniques of randomizing the quantizer breakpoints as well as a more general technique of randomization called dithering were introduced and defined. These techniques were applied to all three of the above quantizers and their effects upon the ARE were observed.

The results of these calculations were presented in either graphs or tables.

6. CONCLUSION

It is difficult to draw any general conclusions from these results. It appears that the decision on whether or not to use a quantizer must be based on an analysis of the problem at hand. The regions where good performance is achieved are usually quite narrow and this reflects a necessarily more detailed knowledge of the noise than may be available when one is concerned with robust systems. At the other extreme, good performance is frequently achieved for very high values of ϵ, at which point nonparametric systems might be better applied.

The possible exception to this general trend is the nonrandom quantizer as a replacement for the limiter in a limiter-correlator detector. Here it appears that there may be a large class of mixture models where a low level number quantizer easily outperforms the limiter. Again it is necessary to consider each model separately since no rule of thumb is obvious.

Finally, the one uniform result of this work has been that neither randomization nor dithering significantly improves the efficacy of any of these quantizers. In fact, the performance is usually far worse. As was stated in the Introduction, the motivation for testing these techniques was the success found in applying similar methods to picture processing. The difference is in the way that the output is produced. The decision system must of course produce a hard decision - H_0 or H_1 - and its performance is measurable in hard numbers - probability of error, false alarm rate, etc. A video signal on the other hand produces an output which is further processed by the human mind and is subject to many psychological and physical effects.

The difference between the two processes seem great enough to state that no real comparison is possible, at least without an accurate model of the human visual process.

In short then, as far as this work has shown, we may conclude that the techniques of randomization and dithering appear to offer little in the way of improving quantized decision systems.

APPENDIX A

Derivation of the Expression of the Efficacy of a Quantizer

The following development applies to the use of a quantizer as illustrated in Fig. 2 in a detection system similar to Fig. 1.

The system produces the statistic $T(x_i)$ given below:

$$T(x_i) = Q(x_i) \tag{A.1}$$

where

$$Q(x_i) = q_k \; ; \qquad x_i \in [t_{k-1}, t_k) \tag{A.2}$$

From (1.2) we then obtain the general expression for the efficacy

$$E_Q = \lim_{n\to\infty} \frac{\left\{ \left(\partial E\left[\sum_{i=1}^{n} s_i Q(x_i) \middle| H_1\right]/\partial\theta\right)\Big|_{\theta=0} \right\}^2}{n \operatorname{Var}\left[\sum_{i=1}^{n} s_i Q(x_i) \middle| H_0\right]} \tag{A.3}$$

As in the preceding work we will let $f(x)$ and $F(x)$ represent the probability density function and distribution function respectively. Assume that $f(-x) = f(x)$, (i.e. $f(x)$ is an even function).

We will now use (A.2) and evaluate the numerator of (A.3)

Since the received signal has sample-wise independent values:

$$E\left[\sum_{i=1}^{n} s_i Q(x_i) \middle| H_1\right] = \sum_{i=1}^{n} s_i E[Q(x_i)|H_1] = \sum_{i=1}^{n} s_i E[Q(n_i + \theta s_i)]$$

$$= \sum_{i=1}^{n} s_i \cdot \sum_{k=1}^{m} \left[q_k\{F(t_k - \theta s_i) - F(t_{k-1} - \theta s_i)\} - q_k\{F(-t_{k-1} - \theta s_i) - F(-t_k - \theta s_i)\}\right]$$

where $t_m \triangleq +\infty$, $-t_m \triangleq -\infty$.

Now

$$\left(\partial E\left[\sum_{i=1}^{n} s_i Q(x_i)\,\middle|\,H_1\right]/\partial\theta\right)\Big|_{\theta=0} =$$

$$\sum_{i=1}^{n} s_i \cdot \sum_{k=1}^{m} s_i\left[-q_k\{f(t_k-\theta s_i)-f(t_{k-1}-\theta s_i)\}\right.$$

$$\left.+\, q_k\{f(-t_{k-1}-\theta s_i)-f(-t_k-\theta s_i)\}\right]\Big|_{\theta=0}$$

$$= 2\sum_{i=1}^{n} s_i^2 \sum_{k=1}^{m} q_k[f(t_{k-1}) - f(t_k)] \qquad (A.4)$$

Similarly we find the variance under H_0:

$$\mathrm{Var}\left[\sum_{i=1}^{n} s_i Q(x_i)\,\middle|\,H_0\right] = \mathrm{Var}\left[\sum_{i=1}^{n} s_i Q(n_i)\right]$$

Again using sample-wise independence and noting that $E[Q(n_i)] = 0$ we have:

$$\mathrm{Var}\left[\sum_{i=1}^{n} s_i Q(n_i)\right] = \sum_{i=1}^{n} s_i^2 E[Q^2(n_i)] \qquad (A.5)$$

Now

$$\sum_{i=1}^{n} s_i^2 E[Q^2(n_i)] = \sum_{i=1}^{n} s_i^2 \sum_{k=1}^{m} [q_k^2\{F(t_k)-F(t_{k-1})\} + q_k^2\{F(-t_{k-1})-F(-t_k)\}]$$

and noting $F(-x) = 1-F(x)$

$$\mathrm{Var}\left[\sum_{i=1}^{n} s_i Q(x_i)\,\middle|\,H_0\right] = 2\sum_{i=1}^{n} s_i^2 \sum_{k=1}^{m} q_k^2[F(t_k) - F(t_{k-1})] \qquad (A.6)$$

Using (A.4) and (A.6) in (A.3) yields the efficacy.

$$E_Q = \lim_{n\to\infty} \frac{4\left(\sum_{i=1}^{n} s_i^2\right)^2 \left\{\sum_{k=1}^{m} q_k[f(t_{k-1}) - f(t_k)]\right\}^2}{2n \sum_{i=1}^{n} s_i^2 \sum_{k=1}^{m} q_k^2[F(t_k) - F(t_{k-1})]} \tag{A.7}$$

If we again let $s_i = 1$ and take the limit, we obtain the expression for the normalized efficacy.

$$\eta_Q = \frac{2\left\{\sum_{k=1}^{m} q_k[f(t_{k-1}) - f(t_k)]\right\}^2}{\sum_{k=1}^{m} q_k^2[F(t_k) - F(t_{k-1})]} \tag{A.8}$$

where $t_m \triangleq \infty$.

If we assume the quantizer is uniform such that $q_k = (2k-1)q_1 = (2k-1)q$ and $t_k = kt$ then

$$\eta_Q = \frac{2\left\{\sum_{k=1}^{m} (2k-1)[f((k-1)t) - f(kt)]\right\}^2}{\sum_{k=1}^{m} (2k-1)^2[F(kt) - F((k-1)t)]} \tag{A.9}$$

APPENDIX B

Derivation of the Expression of the Efficacy of a Randomized Breakpoint Quantizer

Here we will develop an expression similar to (A.8) for the quantizer of Fig. 2 where the breakpoints are a function of a random variable with some probability density g(x) and distribution G(x).

Specifically, we consider all breakpoints t_k, k = 1,...,m, to be functions of a single variable t such that $t_k = c_k t$ where $\{c_k, k = 1,...,m\}$ is a set of m positive constants. For the uniform quantizer, $c_k = k$ and $t = t_1$. Since the results of this paper concern only the uniform quantizer, we consider only this case in our development here.

Proceeding as in Appendix A, using (1.2), (A.1) and (A.2) we can derive an expression similar to (A.3).*

$$E_{RQ} = \lim_{n\to\infty} \frac{\{\partial \int_{-\infty}^{\infty} E[\sum_{i=1}^{n} s_i Q(x_i)|H_1,t]dG(t)/\partial\theta\}^2}{n\int_{-\infty}^{\infty} Var[\sum_{i=1}^{n} s_i Q(x_i)|H_0,t]dG(t)} \tag{B.1}$$

Assuming the necessary conditions, we take the derivative operation inside the integral and proceed as in Appendix A to arrive at the expression for the efficacy

$$E_{RQ} = \lim_{n\to\infty} \frac{4(\sum_{i=1}^{n} s_i^2)^2\{\sum_{k=1}^{m} q_k \int_{-\infty}^{\infty} [f((k-1)t) - f(kt)]dG(t)\}^2}{2n\sum_{i=1}^{n} s_i^2 \sum_{k=1}^{m} q_k^2 \int_{-\infty}^{\infty} [F(kt) - F((k-1)t)]dG(t)} \tag{B.2}$$

where $f(mt) \triangleq 0$, $F(mt) \triangleq 1$.

---

*Note that the sample-wise independence of the random breakpoints has been incorporated implicitly in (B.1).

The normalized expression is then given by

$$\eta_{RQ} = \frac{2\{\sum_{k=1}^{m} q_k \int_{-\infty}^{\infty} [f((k-1)t) - f(kt)]dG(t)\}^2}{\sum_{k=1}^{m} q_k^2 \int_{-\infty}^{\infty} [F(kt) - F((k-1)t)]dG(t)} \tag{B.3}$$

Furthermore, since we are referring to the uniform quantizer, $q_k = (2k-1)q_1 \triangleq (2k-1)q$. Using this in (B.3) we get a final, normalized expression for the efficacy in which the output level q does not appear

$$\eta_{RQ} = \frac{2\{\sum_{k=1}^{m} (2k-1) \int_{-\infty}^{\infty} [f((k-1)t) - f(kt)]dG(t)\}^2}{\sum_{k=1}^{m} (2k-1)^2 \int_{-\infty}^{\infty} [F(kt) - F((k-1)t)]dG(t)} \tag{B.4}$$

APPENDIX C

Derivation of the Expression for the Efficacy of a Dithered Quantizer

Herein we consider the dithered quantizer as illustrated by Fig. 3. The dither signal d_i is considered to be a random variable, the value of which is independent from sample to sample. The probability density and distribution functions are given by $g(x)$ and $G(x)$ respectively.

The test statistic is given by

$$T(x_i) = Q(x_i + d_i) - d_i \tag{C.1}$$

where $Q(\cdot)$ is described by (A.2).

We will consider only the case of uniform quantization so that $q_k = (2k-1)q$ and $t_k = kt$, $t_0 \triangleq 0$. Then (1.2) becomes

$$E_{DQ} = \lim_{n\to\infty} \frac{\left\{ \left(\partial E\left[\sum_{i=1}^{n} s_i (Q(x_i + d_i) - d_i) \middle| H_1 \right] / \partial\theta\right) \Big|_{\theta=0} \right\}^2}{n \operatorname{Var}\left[\sum_{i=1}^{n} s_i (Q(x_i + d_i) - d_i) \middle| H_0 \right]} \tag{C.2}$$

Following Appendix A, we consider first the numerator of (C.2)

$$E\left[\sum_{i=1}^{n} s_i (Q(x_i + d_i) - d_i) \middle| H_1 \right] = \sum_{i=1}^{n} s_i E[Q(n_i + d_i + \theta s_i) - d_i]$$

In all cases considered here, $E[d_i] = 0$, therefore we have under this assumption

$$E\left[\sum_{i=1}^{n} s_i (Q(x_i + d_i) - d_i) \middle| H_1 \right] = \sum_{i=1}^{n} s_i E[Q(n_i + d_i + \theta s_i)] \tag{C.3}$$

We now define $z_i = n_i + d_i$ and make the further assumption that $g(x)$ is an even function, symmetric about $x = 0$. Then, with the same noise density f as in Appendix A, we can find the density $\bar{f}$ of the random variable z_i.

$$\bar{f}(z_i) = \int_{-\infty}^{\infty} f(n_i)g(z_i - n_i)dn_i \qquad (C.4)$$

Let $\bar{F}(z_i)$ be the corresponding distribution function and note that $\bar{f}(z_i)$ is an even, symmetric function. Now (C.3) becomes

$$E\Big[\sum_{i=1}^{n} s_i(Q(x_i + d_i) - d_i)\Big|H_1\Big] = \sum_{i=1}^{n} s_i \sum_{k=1}^{m} \big[q_k\{\bar{F}(t_k - \theta s_i) - \bar{F}(t_{k-1} - \theta s_i)\}$$

$$- q_k\{\bar{F}(-t_{k-1} - \theta s_i) - \bar{F}(-t_k - \theta s_i)\}\big]$$

$$= q \sum_{i=1}^{n} s_i \cdot \sum_{k=1}^{m} (2k-1)\big[\bar{F}(kt - \theta s_i) - \bar{F}((k-1)t) - \theta s_i)$$

$$- \bar{F}-(k-1)t - \theta s_i) + \bar{F}(-kt - \theta s_i)\big]$$

where $\bar{F}(mt) \triangleq 1$.

Then

$$\big(\partial E\big[\sum_{i=1}^{n} s_i T(x_i)\big|H_1\big]/\partial\theta\big)\Big|_{\theta=0} = q \sum_{i=1}^{n} s_i^2 \sum_{k=1}^{m} (2k-1)\big[\bar{f}((k-1)t - \theta s_i) - \bar{f}(kt - \theta s_i)$$

$$- \bar{f}(-kt - \theta s_i) + \bar{f}(-(k-1)t - \theta s_i)\big]\Big|_{\theta=0}$$

$$= 2q \sum_{i=1}^{n} s_i^2 \sum_{k=1}^{m} (2k-1)\big[\bar{f}((k-1)t) - \bar{f}(kt)\big] \qquad (C.5)$$

where $\bar{f}(mt) \triangleq 0$.

Next we consider the variance term,

$$\mathrm{Var}\Big[\sum_{i=1}^{n} s_i T(x_i)\,\Big|\,H_0\Big] = E\Big[\Big(\sum_{i=1}^{n} s_i\{Q(n_i+d_i)-d_i\}\Big)^2\Big] - E^2\Big[\sum_{i=1}^{n} s_i\{Q(n_i+d_i)-d_i\}\Big]$$

$$= E\Big[\sum_{i=1}^{n} s_i^2\{Q(n_i+d_i)-d_i\}^2 + 2\sum_{i=1}^{n-1}\sum_{\ell=i+1}^{n} s_i s_\ell\{Q(n_i+d_i)-d_i\}\{Q(n_\ell+d_\ell)-d_\ell\}\Big]$$

$$- \sum_{i=1}^{n} s_i^2 E^2[Q(n_i+d_i)] - 2\sum_{i=1}^{n-1}\sum_{\ell=i+1}^{n} s_i s_\ell E[Q(n_i+d_i)]\cdot E[n_\ell+d_\ell]$$

and, because of independence between samples,

$$= \sum_{i=1}^{n} s_i^2 E[Q^2(n_i+d_i) - 2d_i Q(n_i+d_i) + d_i^2] + 2\sum_{i=1}^{n-1}\sum_{\ell=i+1}^{n} s_i s_\ell E[Q(n_i+d_i)]E[Q(n_\ell+d_\ell)]$$

$$- \sum_{i=1}^{n} s_i^2 E^2[Q(n_i+d_i)] - 2\sum_{i=1}^{n-1}\sum_{\ell=i+1}^{n} s_i s_\ell E[Q(n_i+d_i)]E[Q(n_\ell+d_\ell)]$$

$$= \sum_{i=1}^{n} s_i^2\{E[Q^2(n_i+d_i)] - E^2[Q(n_i+d_i)] - 2E[d_i Q(n_i+d_i)] + E[d_i^2]\}$$

$$= \sum_{i=1}^{n} s_i^2\{E[Q^2(z_i)] - E^2 Q(z_i)] - 2E[d_i Q(n_i+d_i)] + E[d_i^2]\} \qquad \text{(C.6)}$$

Recalling that $E[d_i] = 0$, (C.6) may be simplified:

$$\mathrm{Var}\Big[\sum_{i=1}^{n} s_i T(x_i)\,\Big|\,H_0\Big] = \sum_{i=1}^{n} s_i^2\{\mathrm{Var}[Q(z_i)] + \mathrm{Var}[d_i] - 2E[d_i Q(n_i+d_i)]\} \qquad \text{(C.7)}$$

where, similarly to (A.6)

$$\mathrm{Var}[Q(z_i)] = 2q\sum_{k=1}^{m}(2k-1)^2[\bar{F}(kt) - \bar{F}((k-1)t)] \qquad \text{(C.8)}$$

and

$$\mathrm{Var}[d_i] = \int_{-\infty}^{\infty} d_i^2\, dG(d_i) \qquad \text{(C.9)}$$

The last term of (C.7) can be evaluated as follows:

$$E[d_i Q(n_i + d_i)] = \int_{-\infty}^{\infty} d_i E[Q(n_i + d_i)|d_i] dG(d_i)$$

$$= \int_{-\infty}^{\infty} d_i \cdot \sum_{k=1}^{m} [q_k\{F(kt-d_i) - F((k-1)t-d_i)\}$$

$$- q_k\{F(-(k-1)t-d_i) - F(-kt-d_i)\}] dG(d_i) \qquad (C.10)$$

Finally the efficacy, using (C.7), (C.5), and (C.2) is

$$E_{DQ} = \lim_{n\to\infty} \frac{4q^2 (\sum_{i=1}^{n} s_i^2)\{\sum_{k=1}^{m} (2k-1)[\bar{f}((k-1)t) - \bar{f}(kt)]\}^2}{\sum_{i=1}^{n} s_i^2\{Var[Q(z_i)] + Var[d_i] - 2E[d_i Q(n_i + d_i)]\}} \qquad (C.11)$$

and, normalized as before with unit signal,

$$\eta_{DQ} = \frac{4q^2\{\sum_{k=1}^{m} (2k-1)[\bar{f}((k-1)t) - \bar{f}(kt)]\}^2}{Var[Q(z_i)] + Var[d_i] - 2E[d_i Q(n_i + d_i)]}$$

REFERENCES

1. P. Huber, "A robust version of the probability ratio test," Annals of Mathematical Statistics, vol. 36, no. 6, December 1965, pp. 1753-1758.

2. R. D. Martin and S. C. Schwartz, "Robust detection of a known signal in nearly Gaussian noise," IEEE Trans. Information Theory, vol. IT-17, no. 1, January 1971, pp. 50-56.

3. J. Capon, "On the asymptotic efficiency of locally optimum detectors," IRE Trans. Information Theory, vol. IT-7, April 1961, pp. 67-71.

4. H. V. Poor and J. B. Thomas, "The design of robust quantizer-detectors for signals in additive, contaminated noise," Proceedings of the Tenth Annual Asilomar Conference on Circuits, Systems and Computers, 1976, pp. 83-87.

5. H. V. Poor and J. B. Thomas, "Asymptotically robust quantization for detection," IEEE Trans. Information Theory, vol. IT-24, no. 2, March 1978, pp. 222-229.

6. S. A. Kassam, "Optimum quantization for signal detection," IEEE Trans. on Communications, vol. COM-25, no. 5, May 1977, pp. 479-484.

7. J. Max, "Quantizing for minimum distortion," IRE Trans. on Information Theory, vol. IT-6, March 1960, pp. 7-12.

8. L. G. Roberts, "Picture coding using pseudo-random noise," IRE Trans. on Information Theory, vol. IT-18, February 1962, pp. 145-154.

9. J. E. Thompson and J. J. Sparkes, "A pseudo-random quantizer for television signals," Proceedings of the IEEE, vol. 55, no. 3, March 1967, pp. 353-355.

10. S. A. Kassam and J. B. Thomas, "Dead-zone limiter: An application of conditional tests in nonparametric detection," J. Acoust. Soc. Am., vol. 60, no. 4, October 1976, pp. 857-862.

11. A. Gersho, "Principles of quantization," IEEE Trans. on Circuits and Systems, vol. CAS-25, vol. 7, July 1978, pp. 427-436.

12. J. D. Bruce, "Optimum quantization for a general error criterion," Quarterly Progress Report No. 69, Research Laboratory of Electronics, M. I. T., pp. 135-141.

13. D. K. Sharma, "Design of absolutely optimal quantizers for a wide class of distortion measures," IEEE Trans. on Information Theory, vol. IT-24, no. 6, November 1978, pp. 693-702.

14. H. V. Poor and J. B. Thomas, "Applications of Ali-Silvey distance measures in the design of generalized quantizers for binary decision systems," IEEE Trans. on Communications, vol. COM-25, no. 9, September 1977, pp. 893-900.

15. J. H. Miller and J. B. Thomas, "Detectors for discrete-time signals in non-Guassian noise," IEEE Trans. on Information Theory, vol. IT-18, March 1972, pp. 241-250.